D06

ADVANCING DATA MANAGEMENT IN MOUNTAIN HAZARD RESEARCH: STRATEGIES FOR ENSURING DATA QUALITY AND ENHANCING MODELING CAPABILITIES

« BACK TO CONTENT

Laura Waltersdorfer; AUSTRIA
Andrea Siposova; Matthias Schlögl; Rudolf Mayer

ABSTRACT

Natural hazards constitute an omnipresent threat in mountainous regions such as the Austrian Alps. Over time, this threat has ushered a shift from heuristic hazard reduction strategies to a quantified risk culture. However, the concept of a quantitative risk assessment essentially builds upon knowledge about the frequency of the considered hazard processes. Thus, information about past events and their properties is a crucial cornerstone for quantitative risk assessment, as it determines the performance and general applicability of methodological toolkits used. Here we present challenges encountered, findings made and lessons learnt from a research project focusing on data-driven susceptibility assessment for shallow landslides in two federal states of Austria.

INTRODUCTION

Climate change is taking an increasing toll on both human lives and assets, with losses from natural hazards being on the rise globally. Climate change impacts on mountain areas such as the Austrian Alps are expected to be particularly severe, as these areas are likely to witness above-average changes. Specifically, changes in both frequency and intensity of extreme weather events are expected to result in more frequent gravitational natural hazard events. The high damage potential of such gravitational mass movements underscores the importance of strengthening knowledge about the likelihood of their present and future occurrence.

Within the Austrian research project gAia (Lampert et al. 2022; Siposova et al. 2023) we focus on shallow landslides and seek to pursue a data-driven approach for providing stakeholders with actionable knowledge to increase preparedness, aid decision-making, and support adaptation measures. Naturally, the general applicability and performance of empirical approaches heavily depend on both the quantity and the quality of the available input data sets. This is particularly crucial in the case of landslide inventories, which are used as training labels for a wide range of machine learning (ML) tasks in this context (Steger et al. 2016). Ideally, such inventories should fulfill certain quality criteria in terms of unbiasedness, completeness, robustness and accuracy. However, these requirements are often not met in practice due to a wide variety of reasons and the limitations of such data are also not adequately reported in terms of metadata or other formal data management standards. Overall, this creates substantial obstacles for modeling and forecasting capabilities.
In this paper, we present data management best practices and tools designed to tackle these challenges and improve the quality of event inventories. We emphasize that the quality of the inventories directly influences the reliability of the outcomes ("garbage in, garbage out"). Hence, addressing these challenges at the foundational level of data collection is critical for ensuring the effectiveness of data-driven approaches.

We recommend refining the current practices by adopting FAIR Principles (Findability, Accessibility, Interoperability, Reusability) (Wilkinson et al. 2016) principles as guiding standards for data management strategies, along with the development of Data Management Plans (DMPs) to systematically organize and oversee these approaches. Furthermore, we advocate for the automation of logging and documentation. This is aimed at ensuring that the inventories are comprehensive and consistent, thereby facilitating their utility in high-quality machine learning applications.

The remainder of this article is structured as follows: We shortly discuss encountered challenges and introduce a conceptual framework (Sec. Methods) synthesizing best practices from FAIR principles, provenance and traceability. We demonstrate the applicability of this method in a case study in context of the gAIa project, exemplifying selected proposed steps (Sec. Discussion) and in- and outputs in a geospatial setting.

METHODS

As previously mentioned, inventories should ideally fulfill quality criteria in terms of accuracy, completeness and statistical representativeness, but may in reality lack certain characteristics because of multiple challenges. Effects such as, e.g., underreporting of smaller events in earlier years, inconsistencies introduced by different modi operandi of documentation and different case workers, lack of information on the exact timestamp of event occurrence as well as uncertainties concerning the spatial location of events lead to incomplete and inconsistent inventories that may constitute a fraught data basis characterized by potentially severe biases. This is further aggravated by the lack of formal data management standards.

Best practices from areas such as FAIR principles, (meta)-data management, provenance and traceability research can be taken, but need to be integrated and adapted to the field. Without taking into account input data quality, complex forecasting methods (e.g. Artificial Intelligence) cannot be robustly applied to natural hazards assessment.

Thus, we propose four actionable steps to improve data management, focusing on data properties and processes (producing data) in particular. We built upon established concepts of traceability (Cleland-Huang et al. 2014) and provenance (Miles et al. 2011), while connecting each process step to state of the art methods and tools of FAIR data management and data-centric artificial intelligence (cf. Fig. 1):

Fig. 1: Overview of the conceptual data management workflow and the associated inputs and outputs corresponding to the single process steps.

1. Identification of data sources and contents: The first step is focused on identifying data sources and formats. In scientific projects, this information can be collected in a (machine-readable) Data Management Plan (ma-DMP) (Miksa et al. 2023) to obtain an overview of all data sources and most important characteristics (e.g. format, size, license, context). This way, project stakeholders have an overview of all relevant data sources and a minimum set of information. This step should also help to identify existing metadata of reused data. This information can be then used and be compared against in subsequent steps.
2. Definition of processing activities: To make results more reproducible and trustworthy, all data producing steps should be made explicit by defining a workflow describing processing activities, as well as the input and output variables (cf. Step 1 to check for consistency between the data management plan and the actual workflow). Established standards for representing complex workflows, such as Business Process Management (BPMN) (White 2004) or semantic web technologies (PROV-DM (PROV-DM), P-Plan Ontology (The P-Plan Ontology)) should and can be reused and adapted.
3. Definition of (meta)-data and process activities trace templates: For all datasets, a metadata template should be provided containing the most relevant characteristics. Furthermore, for all relevant processing activities, provenance traces should be collected to increase the transparency of processing and enable monitoring of the process. Based on the purpose for the documentation, state-of-the-art approaches can be reused and adapted for this purpose: Datasets can be documented using, e.g., datasheets (Gebru et al. 2021). For machine learning models, this includes, e.g., model cards (Mitchell et al. 2019), as well as ML experiment tracking and model registry tools (e.g., MLflow (https://mlflow.org/), Weights & Biases (https://wandb.ai/). The overall workflow can be formalized using task orchestration platforms for data engineering pipelines ((e.g., Kubeflow (https://www.kubeflow.org/), Airflow (https://airflow.apache.org/)). Domain-specific vocabularies (Hungr et al. 2014; Themessl et al. 2022, UNISDR 2009) can further increase the interoperability of the produced results to further increase the reusability.
4. Monitoring processes for natural hazard event data: After defining data sources and trace templates for both data and processes, monitoring processes should be implemented to make sure that the processing activities, and data adheres to certain defined quality metrics (Heinrich et al. 2018; Themessl et al. 2022). This can range from predefined thresholds for certain values, timelessness values or more complicated, qualitative checks on aggregated values. After the satisfaction of internal goals, results should be published and made available to others in machine-readable formats and with rich documentation, one approach to bundle the outputs is RO-Crate (Soiland-Reyes et al. 2022).

Wherever feasible, processing steps are conducted programmatically in Python, R and occasional shell scripts. We use Git for version control and GitHub (https://github.com/) as a tool for facilitating collaboration and structuring technical tasks. In addition, we strive to adhere to good practices in scientific computing (Wilson et al. 2017).

While these steps are generic data management steps, we show their applicability to the natural hazard research context in the next section.

DISCUSSION

To exemplify our method, we present the application of our method to the context of the gAia project. We iteratively developed the method during the project to support reviewability of processes and data based on literature research and feedback from project partners. Project members found the visualization of main processes helpful for discussions, integration points and further enriched it with information concerning their actions.

We model crucial data processing activities (prov:activities) as a P-Plan instantiation (Fig. 2 and Fig. 3), representing the output of Step 2 - Definition of Processing Activities (cf. Fig. 1) and briefly discuss them:

The gAia workflow consists of four major phases: (i) Input preparation, (ii) pre-modeling, (iii) deep learning model training and processing and (iv) output generation. Each of the phases is then specified in concrete activities, for which parameters and metadata are tracked to increase provenance and reviewability. Example metadata include timestamps, creators and file size. We briefly describe more specific metadata for single steps. Input and output data of each activity should be already specified in the Data Management Plan from the first step (cf.

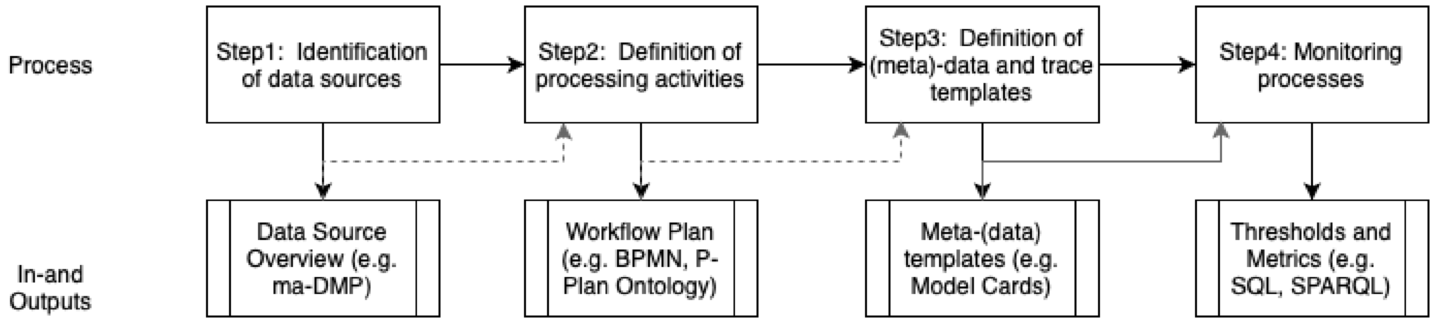
Process
Step1: Identification of data sources
Step2: Definition of processing activities
Step3: Definition of (meta)-data and trace templates
Step4: Monitoring processes
In-and Outputs
Data Source Overview (e.g. ma-DMP)
Workflow Plan (e.g. BPMN, P-Plan Ontology)
Meta-(data) templates (e.g. Model Cards)
Thresholds and Metrics (e.g. SQL, SPARQL)

Fig. 1) and can be mapped into the workflow manually or automatically.

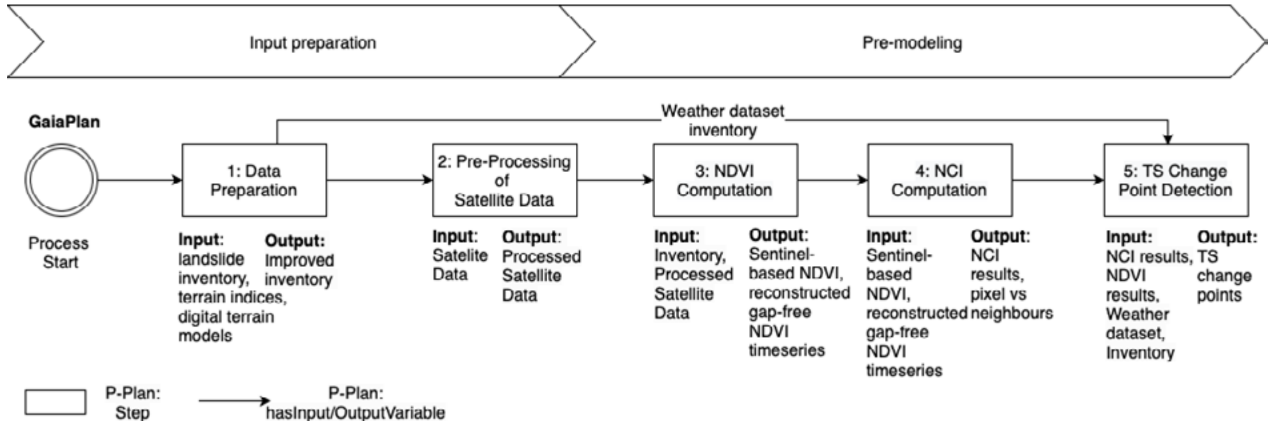
Input preparation
Pre-modeling
GaiaPlan
Process Start
Weather dataset inventory
1: Data Preparation
2: Pre-Processing of Satellite Data
3: NDVI Computation
4: NCI Computation
5: TS Change Point Detection
Input: landslide inventory, terrain indices, digital terrain models
Output: Improved inventory
Input: Satelite Data
Output: Processed Satellite Data
Input: Inventory, Processed Satellite Data
Output: Sentinel-based NDVI, reconstructed gap-free NDVI timeseries
Input: Sentinel-based NDVI, reconstructed gap-free NDVI timeseries
Output: NCI results, pixel vs neighbours
Input: NCI results, NDVI results, Weather dataset, Inventory
Output: TS change points
P-Plan: Step
P-Plan: hasInput/OutputVariable

Fig 2: gAia Workflow Part I: Phases Data Preparation and Pre-Modeling

(i) The first phase, input preparation, consists of (1) data preparation and (2) pre-processing of satellite data. Data preparation: Data to be used in various modeling steps, such as, e.g., landslide inventories from different sources, digital terrain models, geological maps or climate data, is collected and pre-processed (e.g. reprojection to a consistent coordinate reference system, computation of terrain indices and climate indices). Pre-processing of satellite data: Important preprocessing and transformation steps, such as applying atmospheric correction, performing georeferencing, mosaicing, unification and applying cloud masks, are carried out. The Normalized Difference Vegetation Index (NDVI) is computed from the preprocessed satellite images and used as an input for the subsequent steps. Beyond traditional provenance, other metadata such as data sources, limitations of collected data and processing parameters and outputs need to be tracked to support quality estimations of the provided data and the robustness of model outcomes.

(ii) The second phase, pre-modeling, consists of Neighborhood Correlation Image (NCI) (Im and Jensen 2005) computation and time-series (TS) change point detection. NCI computation is utilized as a technique to derive additional features from the NDVI images in order to incorporate the information of pixel neighborhood. Important task-specific characteristics to track would be for example the neighborhood size, which impacts the overall computation. In TS change point detection, the newly derived features are then, in conjunction with the NDVI images, used to determine the amount of change for each pixel. The change scores are aggregated by determining their respective weight, based on their contribution to the event detection.

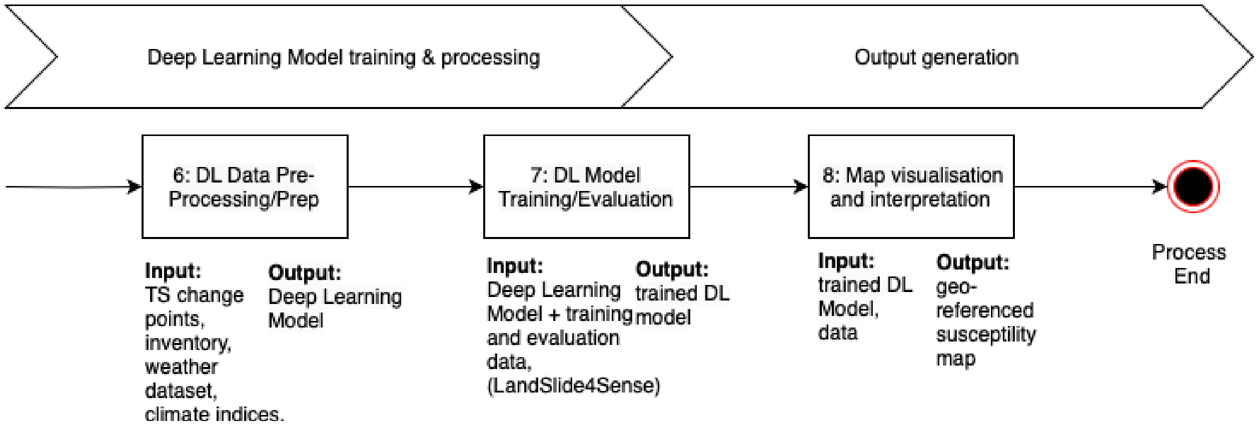
Deep Learning Model training & processing
Output generation
6: DL Data Pre-Processing/Prep
7: DL Model Training/Evaluation
8: Map visualisation and interpretation
Process End
Input: TS change points, inventory, weather dataset, climate indices.
Output: Deep Learning Model
Input: Deep Learning Model + training and evaluation data, (LandSlide4Sense)
Output: trained DL model
Input: trained DL Model, data
Output: geo-referenced susceptility map

Fig 3: gAia Workflow Part II: Phases Deep Learning Model training & processing and Output generation

(iii) The third phase, deep learning model training and processing, consists of additional data preparation (i.e., preparing data for being ingested in deep learning modeling frameworks), as well as model training, validation and evaluation. We leverage repeated nested spatial resampling for model performance estimation. Training data and hyperparameters need to be collected to ensure reproducibility of this step. Relevant metadata (e.g. intended use case, training data, evaluation metrics) for the deep learning models are collected and provided in a Model Card format for both internal and external documentation.

(iv) The final and fourth phase, output generation and interpretation, then consists of producing (1) readily usable geospatial data following common OGC standards (OGC) and (2) providing insights into model interpretability (Molnar 2023). Output data sets are supplemented with appropriate documentation metadata standards and style files for consistent visualization of the generated output in geographic information systems. Additional important information derived by means of machine learning model diagnostics is also provided to facilitate the interpretability of results.

CONCLUSIONS

Data management in general and management of geospatial data in particular can pose major challenges for scientists and practitioners alike. However, proper data management and the adherence to FAIR data principles greatly facilitates data-driven research as well as general usability of data in a practical context.

Throughout this paper, we have identified and examined various challenges posed by current common data management practices concerning the quality assumptions of geospatial data. Without making these quality concerns explicit, the quality of developed models and forecasts is also only questionable.

To alleviate this and bring the metadata of geospatial data more into focus, we have proposed possible solutions to improve selected data management difficulties and discussed approaches to support and automate data management steps. Each step of our framework builds upon each other: Identification of data sources and information collection in a ma-DMP supports findability and accessibility of the data and can act as a basis for provenance and traceability measures throughout the later steps. Definition of processing activities and the resulting workflow formalization support reproducibility by making core activities explicit and linking them back to the data inputs and outputs described in the ma-DMP. The definition of (meta-)data and trace templates based on the workflow formalization is crucial to interoperability and reusability of created and processed data, as well as traceability. Lastly, monitoring processes further ensure the adherence to data quality standards and increase reusability of the produced outputs and models. Each of the four steps of this workflow contribute to multiple aspects of transparent and reliable data management.

Specifically, we have described a machine-readable workflow composed of the most relevant processing activities (Step 2) established in the context of assessing shallow landslide occurrence in Austria. This workflow has already proven beneficial during the project implementation phase for coordinating tasks and activities, but will also be used to communicate main findings to external stakeholders and for future project documentation. Selected relevant metadata for processing activities is discussed to showcase what kind of provenance information is important for different stakeholders throughout the entire process. One limitation so far is the customisation of the provenance model and the initial state of quality levels, which we want to extend and map to existing resources and ontologies. Furthermore, we want to expand the advantages of our approach in the future by providing complex provenance questions for quality assurance and reviewability along the entire workflow process.

By implementing these strategies, we aim to streamline the data management process, enhancing the integrity and reliability of the data. Ultimately, this contributes to the creation of improved inventories which serve as a robust basis for the development of machine learning models in support of disaster risk reduction.

ACKNOWLEDGMENTS

gAia is funded through the KIRAS Security Research Program for Cooperative Research and Innovation Projects by the Austrian Research Promotion Agency (FFG) and the Federal Ministry of Finance, under grant agreement FO999886369 (KIRAS Sicherheitsforschung 2021).

REFERENCES

Cleland-Huang J., Gotel O.C.Z., Huffman Hayes J., Mäder P., Zisman A. (2014) Software traceability: trends and future directions. Future of Software Engineering Proceedings 55–69. https://doi.org/10.1145/2593882.2593891

Gebru T., Morgenstern J., Vecchione B., Vaughan J.W., Wallach H., III H.D., Crawford K. (2021) Datasheets for datasets. Communications of the ACM 64: 86–92. https://doi.org/10.1145/3458723

Heinrich B., Hristova D., Klier M., Schiller A., Szubartowicz M. (2018) Requirements for Data Quality Metrics. Journal of Data and Information Quality 9: 12:1-12:32. https://doi.org/10.1145/3148238

Hungr O., Leroueil S., Picarelli L. (2014) The Varnes classification of landslide types, an update. Landslides 11: 167–194. https://doi.org/10.1007/s10346-013-0436-y

Im J., Jensen J.R. (2005) A change detection model based on neighborhood correlation image analysis and decision tree classification. Remote Sensing of Environment 99: 326–340. https://doi.org/10.1016/j.rse.2005.09.008

KIRAS Sicherheitsforschung Predicting landslides - Development of hazard-warning maps for landslides from consolidated data inventories, https://www.kiras.at/en/financed-proposals/detail/predicting-landslides-entwicklung-von-gefahrenhinweiskarten-fuer-hangrutschungen-aus-konsolidierten-inventardaten

Lampert J., Wernhart S., Avian M., Schlögl M., Seewald M., Jung M., Ostermann M., Kastner R., Mayer R., Siposova A. (2022) gAia: predicting landslides based on consolidated inventory data – bridging needs and limitations. Konferenzband der Disaster Research Days 2022 43–45.

Miksa T., Suchánek M., Slifka J., Knaisl V., Ekaputra F.J., Kovacevic F., Ningtyas A.M., El-Ebshihy A., Pergl R. (2023) Towards a Toolbox for Automated Assessment of Machine-Actionable Data Management Plans. Data Science Journal 22: 1–13. https://doi.org/10.5334/dsj-2023-028

Miles S., Groth P., Munroe S., Moreau L. (2011) PrIMe: A methodology for developing provenance-aware applications. ACM Transactions on Software Engineering and Methodology 20: 8:1-8:42. https://doi.org/10.1145/2000791.2000792

Mitchell M., Wu S., Zaldivar A., Barnes P., Vasserman L., Hutchinson B., Spitzer E., Raji I.D., Gebru T. (2019) Model Cards for Model Reporting. Proceedings of the Conference on Fairness, Accountability, and Transparency 220–229. https://doi.org/10.1145/3287560.3287596

Molnar C. (2023) Interpretable Machine Learning. https://christophm.github.io/interpretable-ml-book/, Leanpub.

OGC, Open Geospatial Consortium Standards, https://www.ogc.org/standards/

PROV-DM The PROV Data Model, https://www.w3.org/TR/prov-dm/

Siposova A., Mayer R., Schlögl M., Lampert J. (2023) Supporting landslide disaster risk reduction using data-driven methods. ERCIM NEWS-European Research Consortium for Informatics and Mathematics 135: 10–11.

Soiland-Reyes S., Sefton P., Crosas M., Castro L.J., Coppens F., Fernández J.M., Garijo D., Grüning B., La Rosa M., Leo S., Ó Carragáin E., Portier M., Trisovic A., RO-Crate Community, Groth P., Goble C. (2022) Packaging research artefacts with RO-Crate. Data Science 5: 97–138. https://doi.org/10.3233/DS-210053

Steger S., Brenning A., Bell R., Glade T. (2016) The propagation of inventory-based positional errors into statistical landslide susceptibility models. Natural Hazards and Earth System Sciences 16: 2729–2745. https://doi.org/10.5194/nhess-16-2729-2016

The P-Plan Ontology, http://vocab.linkeddata.es/p-plan/index.html

Themessl M., Enigl K., Reisenhofer S., Köberl J., Kortschak D., Reichel S., Ostermann M., Kienberger S., Tiede D., Bresch D.N., Röösli T., Lehner D., Schubert C., Pichler A., Leitner M., Balas M. (2022) Collection, Standardization and Attribution of Robust Disaster Event Information—A Demonstrator of a National Event-Based Loss and Damage Database in Austria. Geosciences 12: 283. https://doi.org/10.3390/geosciences12080283

White S.A. (2004) Introduction to BPMN. Ibm Cooperation. 2: 0

Wilkinson M.D., Dumontier M., Aalbersberg Ij.J., Appleton G., Axton M., Baak A., Blomberg N., Boiten J.-W., da Silva Santos L.B., Bourne P.E., Bouwman J., Brookes A.J., Clark T., Crosas M., Dillo I., Dumon O., Edmunds S., Evelo C.T., Finkers R., Gonzalez-Beltran A., Gray A.J.G., Groth P., Goble C., Grethe J.S., Heringa J., 't Hoen P.A.C., Hooft R., Kuhn T., Kok R., Kok J., Lusher S.J., Martone M.E., Mons A., Packer A.L., Persson B., Rocca-Serra P., Roos M., van Schaik R., Sansone S.-A., Schultes E., Sengstag T., Slater T., Strawn G., Swertz M.A., Thompson M., van der Lei J., van Mulligen E., Velterop J., Waagmeester A., Wittenburg P., Wolstencroft K., Zhao J., Mons B. (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3: 160018. https://doi.org/10.1038/sdata.2016.18

Wilson G., Bryan J., Cranston K., Kitzes J., Nederbragt L., Teal T.K. (2017) Good enough practices in scientific computing. PLOS Computational Biology 13: e1005510.

https://doi.org/10.1371/journal.pcbi.1005510

UNISDR terminology on disaster risk reduction | UNDRR, http://www.undrr.org/publication/2009-unisdr-terminology-disaster-risk-reduction