This project has received funding from
the European Union's Horizon Europe
research and innovation programme
under grant agreement No. 101059548.

OPEN EARTH MONITOR

D7.1 AI Code of Conduct and Open-Earth-Monitor Harmonization Guidelines

Document control page

Project	Open-Earth-Monitor (OEMC)
Project, full title	A cyberinfrastructure to accelerate uptake of environmental information and help build user communities at European and global levels
Project number	101059548
Project start	June 1st 2022
Deliverable	DX7.1 Report "AI code of conduct and Open-Earth-Monitor harmonization guidelines"
Work Package	WPX TITLE
Document title	AI code of conduct and Open-Earth-Monitor harmonization guidelines
Version	final
Responsible author	Gilberto Camara

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101059548.

Contributors	Ichsani Wheeler, Tom Henlg, Carson Ross, Igor Milosavljevic
Date of delivery	30th November 2022
Due date of deliverable	30th November 2022
Type	Report
Language	English
Rights	
Status	( ) In progress ( ) For review (x) Approved
Dissemination level	Confidential

Table of contents

Executive summary

1. This document presents the guidelines to be followed by the Open Earth Monitor project to ensure compliance with well-established data access and data sharing principles and the European General Data Protection Regulation. It will also provide guidelines regarding Ethical issues when using artificial intelligence (AI) and machine learning (ML) methods.
2. The different communities that deal with environmental and scientific data have agreed on a number of principles that should guide data access, management and curation.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101059548.

Each of these principles has been established by a specific community. Although they have broadly the same intention, which is ensuring open data access in a way that respects the interests of both producers and users, they differ in some key points.

3. The document discusses different principles for data access and sharing and outlines strategies on how to deal with each one.

General Approach

Since the Open Earth Monitor project will build a single cyberinfrastructure for different applications of Earth observation data, and these applications are built by different project partners, harmonization plays a key role in the project. This document considers different aspects of harmonization and will serve as a guideline for the partners. In what follows, we discuss the following issues:

a) Community-based principles for open data sharing;

b) European frameworks related to data access rights and privacy protection;

c) Technical matters that require specific actions by the project partners;

d) Ethics in AI issues.

The main output of the document is a series of recommendations that will be shared with the project partners and used by the project management team to ensure compliance with the stated principles. It will be the role of Task T7.1 (Implementation of data as empowering governance) to verify such compliance and interact with the partners so that the agreed principles are followed.

FAIR Data Principles

The FAIR principles have been proposed by the scientific community to enable reuse of scholarly data (Wilkinson et al. 2016). The FAIR acronym stands for the aims of **F**indability, **A**ccessibility, **I**nteroperability, and **R**euse of digital assets. These aims are expressed in properties that should be followed by data producers.

For finding data, the following principles apply:

F1. (Meta)data is assigned a globally unique and persistent identifier.
F2. Data are described with rich metadata (defined by R1 below).
F3. Metadata clearly and explicitly include the identifier of the data they describe.
F4. (Meta)data are registered or indexed in a searchable resource.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101059548.

By using the STAC protocol, OEMC can satisfy requirements **F1**, **F3** and **F4** by construction. Support for principle **F2** will be ensured by creating a project-wide glossary that will be used to describe OEMC applications.

For accessing data, the FAIR principles are:

> **A1.** (Meta)data are retrievable by their identifier using a standardised communications protocol.
> **A2.** Metadata are accessible, even when the data are no longer available.

By using the STAC protocol, OEMC can satisfy principle A1. Compliance of principle A2 is the responsibility of the OEMC project, as stated in its implementation plan.

In terms of interoperability, the FAIR principles are:

> **I1**. (Meta)data uses a formal, accessible, shared, and broadly applicable language for knowledge representation.
> **I2**. (Meta)data uses vocabularies that follow FAIR principles.
> **I3**. (Meta)data include qualified references to other (meta)data.

Using the STAC protocol will allow OEMC to adhere to principle **I1**. However, principles **I2** and **I3** are more ambitious in scope, because they require active action for the production of vocabularies that describe data and metadata. Given the scope of the applications on the OEMC project, it is recommended that each application developed in the OEMC project makes a strong effort to produce well-defined vocabularies following the generic OEMC glossary. Doing so will allow OEMC to adhere to principles **I2** and **I3**.

In terms of reusability, the FAIR principles require:

R1. (Meta)data are richly described with a plurality of accurate and relevant attributes.
R1.1. (Meta)data is released with a clear and accessible data usage license.
R1.2. (Meta)data is associated with detailed provenance.
R1.3. (Meta)data meet domain-relevant community standards.

Principle R1 can be satisfied if OEMC applications use a project-wide glossary. The OEMC infrastructure will also ensure that each data set released will comply with principles **R1.1**, **R1.2** and **R1.3**.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101059548.

GEOSS Data Sharing Principles

The GEOSS data sharing principles have been developed by the Group on Earth Observations (GEO) to support full, free and open access to Earth observation data. These principles are:

- data, metadata and products will be shared as Open Data by default;
- where international instruments, national policies or legislation preclude the sharing of data as Open Data, data should be made available with minimal restrictions on use and at no more than the cost of reproduction and distribution; and
- all shared data, products and metadata will be made available with minimum time delay.

To the extent that the OEMC project is committed to build an infrastructure which shares data in a free and open manner, the project adheres fully with the GEOSS Data Sharing Principles.

CARE Principles

The CARE Principles for Indigenous Data Governance were established in the context of the United Nations Declaration on the Rights of Indigenous Peoples. These principles reflect the interest of indigenous communities worldwide. These communities possess considerable knowledge about their lands, having inhabited the same areas for many generations. Many of the areas occupied by indigenous peoples, such as the Arctic, Siberia and Amazonia, have considerable importance in terms of global sustainability and climate change. Traditionally, there has been a disconnection between researchers working in areas occupied by traditional people and these communities. The responsibilities and rights of local communities are not always respected. For this reason, indigenous communities and researchers germane to their rights have promoted a new set of principles, that address concerns related to indigenous communities about data collection on their lands and to the right to fair share of benefits associated with such data (Carroll et al. 2021).

The main pillars of the CARE principles are:

a) Collective benefits;
b) Authority to contr**a**ol;
c) Responsibility and Ethics.

The CARE guidance is that data collection activities in areas associated to indigenous communities should be respectful of those communities and their cultures. One example is data collection associated with biodiversity measures. In many cases, researchers doing

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101059548.

fieldwork have not properly acknowledged direct or indirect contributions of indigenous knowledge in their findings (Carroll et al. 2021).

In relation to the Open Earth Monitor project, the CARE principles are mostly relevant to the use cases involving or related to indigenous communities. Tasks that may involve data associated with indigenous communities include:

T6.1 (Development of the World-forest monitor),
T6.2 (Development of the tropical-deforestation monitor),
T6.3 (Development of the tropical-crop monitor), and
T6.4 (Development of the world-land degradation neutrality monitor).

It will be the responsibility of the OEMC management, and more specifically of Task T7.1 (Implementation of data as empowering governance) to work with the afore-mentioned tasks to verify compliance with the CARE principles.

European General Data Protection Regulation

The EU General Data Protection Regulation (GDPR) contains provisions and requirements related to the processing of personal data of individuals (formally called *data subjects* in the GDPR) who are located in the European Economic Area. Similar data protection regulations have been issued by other countries, including the USA, China, Australia, and Brasil. The World Geospatial Industry Council (WGIC) produced a report that analyzes the impact of the EU GDPR and other national regulations in geospatial data (Desmet 2020). This report lists geospatial applications which are affected by the GDPR, including land ownership and risk assessment.

The OEMC project includes various applications that are impacted by EU GDPR. They include:

T5.02 (EU-in-situ-data tool),
T5.03 (EU reforestation planner),
T5.04 (EU forest management),
T5.05 (EU costal monitor),
T5.06 (EU biodiversity monitor),
T5.07 (EU crop monitor),
T5.09 (EU climate monitor),
T5.10 (EU flood monitor),
T5.11 (EU soil monitor),
T5.12 (EU snow monitor),
T5.13 (EU extreme weather risk monitor),
T5.14 (EU rapid forest disturbance),

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101059548.

T5.15 (EU land based mitigation).

In all these cases, OEMC management will interact with the task leaders to ensure the stipulations of the GDPR are adhered to. All OEMC applications affected by EU GDPR should comply with the principles of **"privacy by design"** and **"privacy by default"**. Any personal data should be treated to guarantee confidentiality, integrity and availability of such data.

OEMC will set up a Data Management Program, as required by GDPR, which will be responsible for setting up the strategy for personal data protection. All project partners that provide geospatial data with personal content will be required to sign Data Protection Agreements with third parties that provide such data to them.

AI Code of Conduct

Algorithms that use machine learning (ML) and artificial intelligence (AI) are an important part of most of the tasks developed by the Open Earth Monitor (OEM) project, both in European and in the global case studies. In these and many other cases where they are used, AI/ML methods are at the core of decision-making and play a key role in producing mapping, monitoring and planning tools. For this reason, the OEMC project needs to adhere to a common set of AI/ML Ethics guidelines.

Part of the basis for the establishment of OEMC AI/ML Ethics guidelines is the proposed European Artificial Intelligence Act (EU AI Act), which aims to align AI applications with EU values and fundamental rights. The approach taken by the EU AI Act is risk-based. The concern of the European regulators is to focus on cases that relate to fundamental rights and individual or collective safety. The Act attempts to identify applications and systems that create an unacceptable or high risk. These applications include cases where AI tools are used for social profiling and decision-making in situations that have lasting effects on individual lives. Applications not explicitly banned or listed as high-risk are self-regulated.

The Centre for European Policy Studies (CEPS), a leading think-tank, has issued an assessment of the EU AI Act (Boguki et al. 2022). Their assessment indicates that the current proposal fails to consider indirect effects of AI-based applications. The environmental area is one such case. Although decisions guided by AI in environmental matters rarely affect individual rights directly, misguided or wrongly-designed application could, in principle, have negative effects on collective rights.

To take one example, consider an AI-based application that monitors reforestation and forest management. Forest-based biomass currently makes up nearly 60% of European renewable

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101059548.

energy sources. In September 2022, the European Parliament introduced a cap on the share of "primary woody biomass" that can be counted as a renewable form of energy. This decision leaves open how to identify "primary woody biomass". Therefore, AI-based algorithms that use remote sensing and in-situ data to characterize the different types of European forests are bound to have an impact on how the EU Parliament decision is to be implemented.

The objective of the OEMC project is to produce complete, consistent, environmental data to the highest level of analysis-/decision-readiness to be used by others for governance purposes. Therefore, its applications are based on the principle of 'trustworthy AI'. To achieve this goal, the project's methods use open source algorithms, open access data and reproducible methods.

We recognise that ML/AI algorithms embody values, assumptions, and purposes, whether their programmers consciously intend them to or not. These biases affect their decision-making, which have the potential to not only reduce general accuracy, but also disproportionately disenfranchise specific groups of stakeholders. **Given there is no guaranteed way of avoiding such biases, the approach taken by the OEMC project is to focus on "Reproducible AI".**

In this vein, a group of leading researchers in AI, frustrated by problems with reliability and reproducibility of experiments using AI/ML, identified many gaps present in the scientific literature of the field (Pineau et al. 2021). These gaps include:

- Lack of access to the same training data / differences in data distribution;
- Misspecification or under-specification of the model or training procedure;
- Lack of availability of the code necessary to run the experiments, or errors in the code;
- Under-specification of the metrics used to report results;
- Improper use of statistics to analyze results, such as claiming significance without proper statistical testing or using the wrong statistic test;
- Selective reporting of results and ignoring the danger of adaptive overfitting;
- Over-claiming of the results, by drawing conclusions that go beyond the evidence presented (e.g. insufficient number of experiments, mismatch between hypothesis & claim).

To improve reproducibility and reduce gaps in reuse and reproducibility of AI applications, the leading group of researchers proposes that reporting of AI-based applications should follow a checklist (Pineau et al. 2021). The most relevant items of that checklist for the OEMC project are:

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101059548.

- ☐ A clear description of the mathematical setting, algorithm, and/or model.
- ☐ A link to a downloadable source code, with specification of all dependencies, including external libraries.
- ☐ A link to a downloadable version of the dataset or simulation environment.
- ☐ An explanation of any data that was excluded, description of any pre-processing step.
- ☐ An explanation of how samples were allocated for training / validation / testing.
- ☐ The range of hyper-parameters considered, method to select the best hyper-parameter configuration, and specification of all hyper-parameters used to generate results.
- ☐ A description of how experiments were run.
- ☐ A clear definition of the specific measure or statistics used to report results.
- ☐ A description of the computing infrastructure used.

The above checklist constitutes the "AI Code of Conduct" to be applied in all AI-based applications of the Open Earth Monitor (OEM) project. This code of conduct will be shared with all OEMC partners. The project management team will ensure all applications delivered to the OEMC cyberinfrastructure adhere to this Code of Conduct.

Data Harmonization

Data harmonization plays an important role in the Open Earth Monitor project and assumes different meanings, depending on the task involved. Considering the project's proposal, one can identify the following types of harmonization:

- Labels associated to in-situ data collection and mapping of land-use and land-cover classes (Tasks T4.06, T4.08, T5.02, T5.03, T5.05, T5.07, T5.10, T5.13, T5.13, T6.01, T6.02, T6.03, T6.04)
- Parameters and measurement ranges associated with biomass and carbon emissions (Tasks T4.03, T6.08).
- Parameters and measurement ranges and values associated with meteorological, oceanographic and climate data (Tasks T4.02, T4.07, T4.09, T5.11, T5.14, T6.05, T6.07, T 6.10).
- Collection and modelling of biodiversity (Tasks T4.04, T5.06).

Each of these types of harmonization require a specific approach to ensure consistency and compatibility between the different OEMC applications. Arguably, the simplest case is that of values associated with meteorological and climate data, since the World Meteorological Organisation (WMO) has spent considerable effort on metadata standards for such data. In this

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101059548.

case, the application developers will be asked to use WMO standards and to document such usage with associated metadata.

In the case of biomass and carbon emissions, the International Panel on Climate Change (IPCC) has spent considerable effort. Therefore, OEMC partners are asked to refer to the "2019 Refinement to the 2006 IPCC Guidelines for National Greenhouse Gas Inventories" produced by the IPCC.

Biodiversity data provides an important challenge, since there is no internationally agreed standard for data description. Since there is no globally agreed taxonomy of species names (Grenié et al. 2022), the OEMC project management will interact with project partners responsible for Tasks T4.04, T5.06 to ensure that the data and applications delivered to the OEMC cyberinfrastructure will be reusable and reproducible.

The most challenging harmonization task concerns applications related to land-use and land-cover (LUC). Despite recent progress on standards for land cover description (Herold et al. 2006), harmonization of LUC data remains an intrinsic difficult problem to solve (Jansen, Groom, and Carrai 2008). Some authors argue that there are inherent differences on land classification which are unsolvable by a single class schema (Camara 2020, Chazdon et al. 2016). Given the unsolved state of the land harmonization problem, it is advisable that the OEMC project adheres strongly to the FAIR principles (see above) to ensure reuse and reproducibility.

Final remarks

This document provides a set of recommendations that will enable the Open Earth Monitor (OEM) project achieve its goals of producing algorithms and applications that are openly available in an open cyberinfrastructure. Based on the analysis of the applicability of the FAIR, CARE, and GEOSS Data Sharing Principles, on the perceived impact of the EU GDPR directive and EU AI Act, and considering the need for Reproducible AI and harmonised data sets, we provide the following recommendations:

1. The FAIR principles should be strictly followed by all OEMC applications;
2. Whenever application, OEMC partners need to be aware of the CARE principles and follow them;
3. OEMC should emphasise in its communication strategy that, by design, it follows the GEOSS Data Sharing Principles;
4. OEMC should adopt an AI Code of Conduct that includes best practices recommended by leading AI researchers;

5. OEMC Data Management Plan will include provisions for compliance with the EU General Data Protection Regulation;
6. OEMC will produce a common glossary of terms that will be used to support data harmonization efforts;
7. Data harmonisation for land use and land cover data should be based on methodical production of metadata that allows reuse and reproducibility;
8. OEMC applications related to biodiversity need to be established by interaction between the project partners.
9. OEMC applications related to biomass and carbon emissions should follow IPCC guidelines for greenhouse gas inventories;
10. OEMC applications related to meteorological and climatological data should follow WMO standards, when applicable.

References

Boguki, Artur, Alex Engler, Clement Perarnaud, and Andrea Renda. 2022. 'The AI Act and Emerging EU Digital Acquis'. CEPS (Centre for European Policy Studies). https://www.ceps.eu/ceps-publications/the-ai-act-and-emerging-eu-digital-acquis/.

Camara, Gilberto. 2020. 'On the Semantics of Big Earth Observation Data for Land Classification'. *Journal of Spatial Information Science* 2020 (20): 21–34. https://doi.org/10.5311/JOSIS.2020.20.645.

Carroll, Stephanie Russo, Edit Herczog, Maui Hudson, Keith Russell, and Shelley Stall. 2021. 'Operationalizing the CARE and FAIR Principles for Indigenous Data Futures'. *Scientific Data* 8 (1): 108. https://doi.org/10.1038/s41597-021-00892-0.

Chazdon, Robin L. and others. 2016. 'When Is a Forest a Forest? Forest Concepts and Definitions in the Era of Forest and Landscape Restoration'. *Ambio* 45 (5): 538–50. https://doi.org/10.1007/s13280-016-0772-y.

Desmet, Arnout. 2020. 'Geospatial Information and Privacy (Report by WGIC)'. World Geospatial Industry Council. https://wgicouncil.org/wp-content/uploads/2020/03/Geospatial-Informaion-and-Privacy-report-Final.pdf.

Grenié, Matthias, Emilio Berti, Juan Carvajal-Quintero, Gala Mona Louise Dädlow, Alban Sagouis, and Marten Winter. n.d. 'Harmonizing Taxon Names in Biodiversity Data: A Review of Tools, Databases and Best Practices'. *Methods in Ecology and Evolution* n/a (n/a). Accessed 29 November 2022. https://doi.org/10.1111/2041-210X.13802.

Herold, M., J. S. Latham, A. Di Gregorio, and C. C. Schmullius. 2006. 'Evolving Standards in Land Cover Characterization'. *Journal of Land Use Science* 1 (2–4): 157–68. https://doi.org/10.1080/17474230601079316.

Jansen, Louisa J.M., Geoff Groom, and Giancarlo Carrai. 2008. 'Land-Cover Harmonisation and Semantic Similarity: Some Methodological Issues'. *Journal of Land Use Science* 3 (2–3): 131–60. https://doi.org/10.1080/17474230802332076.

Pineau, Joelle, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Lariviere, and Alina Beygelzimer. 2021. 'Improving Reproducibility in Machine Learning Research'. *Journal of Machine Learning Research* 22: 20.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101059548.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. 'The FAIR Guiding Principles for Scientific Data Management and Stewardship'. *Scientific Data* 3 (1): 160018. https://doi.org/10.1038/sdata.2016.18.