Ref. Ares(2024)3871577 - 30/05/2024

ECHO

ENGAGING CITIZENS IN SOIL SCIENCE:
THE ROAD TO HEALTHIER SOILS

Deliverable 1.3

Citizen-generated Soil Data Quality Assessment Framework

Co-funded by the European Union

UK Research and Innovation

ECHOSOIL.EU

Project information

Project number	101112869
Project acronym	ECHO
Project name	Engaging Citizens in Soil Science: The Road to Healthier Soils
Call	HORIZON-MISS-2022-SOIL-01
Topic	HORIZON-MISS-2022-SOIL-01-09
Type of Action	HORIZON Research and Innovation Actions
Responsible Service	REA.B.2
Project staring date	01 June 2023
Project duration	48 months

Document Details

Deliverable	D1.3 – Citizen-generated Soil Data Quality assessment framework
Work Package	1 – Enabling high-impact citizen science for soil monitoring
Task	1.3 - Citizen-generated Soil Data Quality assessment framework
Deliverable Type	Document, Report
Dissemination Level	Public
Deliverable Lead	FC.ID/CIENCIAS
Date of publication	31/05/2024

Disclaimer

Funded by the European Union under GA no. 101112869 – ECHO and co-funded by UK Research and Innovation (UKRI).

Short description of the deliverable

Deliverable D1.3 – Citizen-Generated Soil Data Quality assessment framework constitutes an integral component of the work conducted within WP1 - Enabling high-impact citizen science for soil monitoring. This deliverable addresses the need of selecting methodologies suitable for utilization within the ECHO project for its citizen science initiatives. The framework will be applied to the projects listed in the matrix created in T1.1 and, together with deliverable D1.2 - Assessment framework for Citizen Science Methods, will enable the evaluation of soil monitoring methods and ensure the quality of the collected data.

Versioning and contribution history

Version	Date	Modified by	Notes
0.1	15/05/2024	Silvana Munzi & Cristina Cruz (FC.ID/CIENCIAS)	Draft version
0.2	21/05/2024	Alba Peiro & Francisco Sanz (IBERCIVIS)	Revised Version 1
0.3	21/05/2024	Tanja Mimmo & Claudia Cappella (UNIBZ)	Revised version 2

Foreword

Soil is a vital, yet often disregarded, resource that supports life on Earth by providing the foundation for agriculture, forests, and various other natural ecosystems. However, soil degradation is a growing concern around the world, and it can have severe consequences for our planet like reduced crop yields, increased greenhouse gas emissions, and decreased biodiversity. The ECHO project aims to prevent this by bringing together citizens and volunteer scientists from around Europe to work towards a common goal of protecting and preserving our soils, thus contributing to the transition towards healthy soils of the EU Mission: "A Soil Deal for Europe".

ECHO will generate new data on the health status of EU soils, complementing existing soil mapping and monitoring in EU Member States and Scotland, including the EU Soil Observatory (EUSO). The project will develop and deploy 28 tailor-made citizen science initiatives across EU Member States and Scotland, taking into account different land-uses, soil types, and biogeographical regions, as well as stakeholder needs. With 16 participants from all over Europe, including 10 leading universities and research centres, 4 SMEs, and 2 Foundations, under the coordination of the Free University of Bolzano-Bozen, ECHO will assess 16,500 sites in different climate and biogeographic regions to achieve its ambitious goals.

The project aims to engage citizens in protecting and restoring soils by building their capacities and enhancing their knowledge. Citizens will thereby not only actively contribute to the project's data collection but also promote soil stewardship and foster behavioural change across the EU. The ECHOREPO, a long-term open access repository with a direct link to the EUSO, will make the citizen science data available for exploitation not only by scientists but also by citizens, policy makers, farmers, landowners and other end-users, providing added value to existing data and other relevant soil monitoring initiatives. ECHOREPO will thus provide valuable information about the state of soil health in various regions, and help citizens make informed decisions about land use and conservation.

We believe that the ECHO project will have a significant impact on soil health and citizen engagement across Europe and become an important step towards protecting and preserving our soil for future generations. By working together, we can ensure that our soil remains healthy and productive, and that we continue to enjoy the many benefits it provides.

Table of Contents

1. Purpose of the assessment framework

One of the main aims of WP1 is identifying citizen science (CS) initiatives focused on soil monitoring, gathering and analyzing pertinent details for ECHO, including methodology, technology employed, scientific data, and engagement strategies. As a result, a review of the State of the Art on CS initiatives for monitoring soil health (soil biodiversity and pollution) in the form of a matrix (Ibercivis Foundation, 2023), during Task 1.1 has been carried out. In this context, the FCiências.ID - Associação para a Investigação e Desenvolvimento de Ciências and the Faculdade de Ciências da Universidade de Lisboa (Portugal) have conducted two assessment frameworks: the "Assessment framework for CS methods for soil monitoring" in Task 1.2, and "Citizen-generated Soil Data Quality Assessment Framework" here in Task 1.3 to evaluate the quality of soil data produced through the methodologies utilized in these projects, which will be included in ECHO's activities.

A Citizen-Generated Soil Data (CGD) quality assessment framework serves as a systematic approach to evaluate and ensure the reliability, accuracy, and usability of soil data. It involves a set of criteria, methods, and procedures designed to assess various aspects of soil data.
Although there are some different definitions (Haklay et al., 2021), CS can be understood as a methodology in which professional scientists engage with non-professional scientists to accomplish scientific research. Due to the wide variability in skills and expertise among contributors, issues of data quality often come to the forefront when considering the validity of CS research (Balázs et al., 2021). Data quality refers to the fitness of data for an intended purpose, and establishing data quality typically involves a multifaceted evaluation of states such as relevance, accuracy, accessibility, coherence, interpretability, comparability, and adaptation to local characteristics.

Providing information about data quality improves opportunities for data reuse by increasing the trustworthiness of the data. Moreover, it helps potential data users to determine whether and how data can be used, facilitating the analysis and interpretation of such data. However, assessing CGD quality can be extremely difficult due to the participation of heterogeneous observers, the variety of methods used and a lack of information about such methods. Data bias, errors, uncertainty, and ethical issues pose challenges that should be assessed regularly as part of the CS research projects. Considering data quality during the project's earliest stages can greatly improve planning and enable the research team to identify issues that might have a subsequent impact on data quality (Wiggins et al., 2011). Information regarding CGD quality should be gathered, stored, and disseminated together with the data to ensure high-quality stewardship.

The challenge of assessing CGD quality is more acute when considering data reuse and alignment with FAIR principles. Building on sound, existing quality assessment frameworks, we propose a list of criteria suitable for CS data that capture both data production quality and usability. Each quality criterion included in the framework is assigned with a score, offering flexibility in rating options to the quality criteria listed and establishing a minimum quality level or threshold for data acceptance.

2. Description of the assessment framework

This deliverable defines the quality assessment framework to leverage the power of CGD, encompassing the following elements: i) a list of quality criteria suitable to evaluate the quality of CGD; ii) a score iii) a minimum quality level or threshold – a number that sets the minimum quality level at which CGD would be acceptable.

The quality assessment framework includes seven dimensions:

1. **Relevance:** degree to which CGD can be used to assess specific soil indicators
2. **Accuracy:** measurements precision
3. **Accessibility:** ease with which CGD are presented, released, and made available to users
4. **Coherence:** comparability over time and across geographical units
5. **Interpretability:** simplicity with which data can be interpreted
6. **Compatibility** of data with the EU Soil Observatory (EUSO) and other existing soil monitoring systems like Land Use and Coverage Area frame Survey (LUCAS), including data formats and classification
7. **Adaptation** to the biogeographical region and soil types: degree to which the method could be replicated to further biogeographical regions and soil types, besides the ones in which it has been implemented.

3. How to use the framework

For each criterion, a set of questions is identified. The scoring system has been designed as a dichotomous path where you can choose one of two answers to each question. One answer allows you to proceed to the next question, where the dichotomy is repeated, until the final answer is reached. The other answer leads to a score or to another question with two possible scores.

The lowest score indicates significant flaws, so serious that even one of them is sufficient to justify the rejection of the methodology. Typically, these situations are also the least likely to occur.

4. Scoring

For each criterion, a scoring scale is established to evaluate its potential and robustness for adoption in this or further research projects. A series of questions takes considers different aspects (sub-criteria) of each criterion, which are scored using the scale in Figure 1.

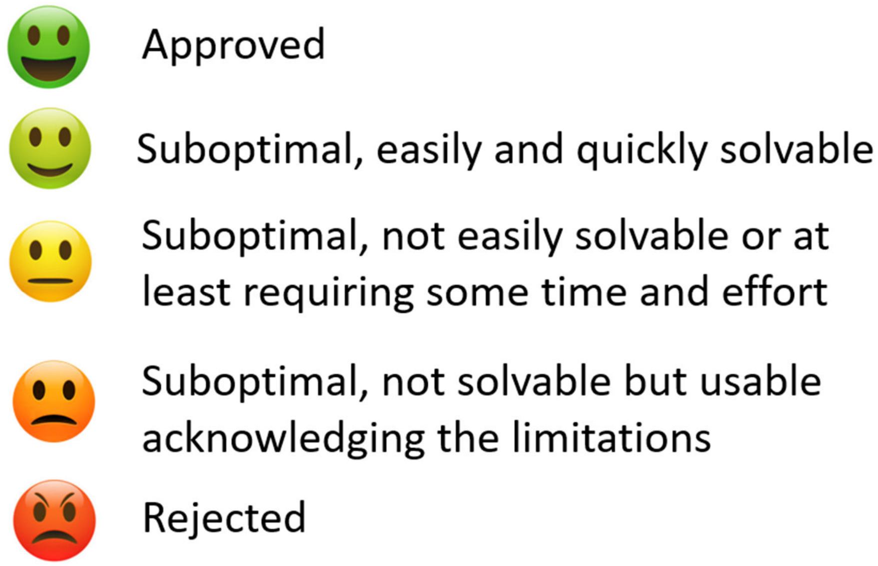

Fig. 1. Possible scores for each sub-criterion considered in the evaluation of CGD.

The establishment of such a scoring scale not only facilitates the systematic evaluation of sub-criteria but also enhances transparency and consistency in the assessment process. By quantifying the potential and robustness of each criterion, researchers can make informed decisions regarding their suitability for integration into the project. Moreover, this approach provides valuable insights into strengths and areas for improvement, thereby guiding future research directions and methodological refinements.

Thresholds (Fig.2) are established based on a comprehensive evaluation of the criteria to determine whether a particular methodology meets the requirements for integration into ECHO's activities.

Fig. 2. Thresholds for the suitability of methodologies for integration into ECHO's activities.

The circles indicate the possible final scores for the methods (Fig. 3).

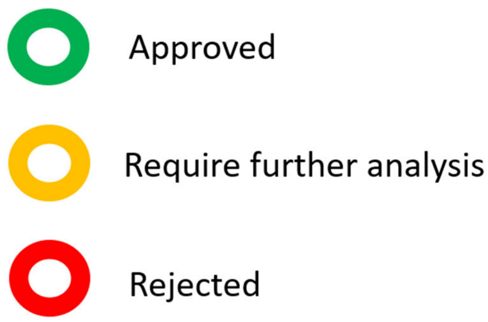

Fig. 3. Final scores for the methodologies to be used in ECHO's activities.

Therefore, beyond mere approval or rejection, a method may require additional analysis and discussion among the partners involved in T2.1 “Selection of citizen science methods for monitoring soils”, allowing a certain degree of flexibility to optimize the selection of soil citizen science methodologies.

5. Dimensions of the assessment framework

5.1. Relevance

This dimension captures the usability of the data to serve a specific purpose sought by the data users, which in this case is to assess specific soil indicators. As implied by various quality assessment frameworks (Balázs et al., 2021), measuring relevance requires the identification of data user groups and their needs. To identify the potential use of CGD on soil indicators, it is essential to identify the data needs of government entities, civil society, academia, the private sector, and citizens (Fig. 4).

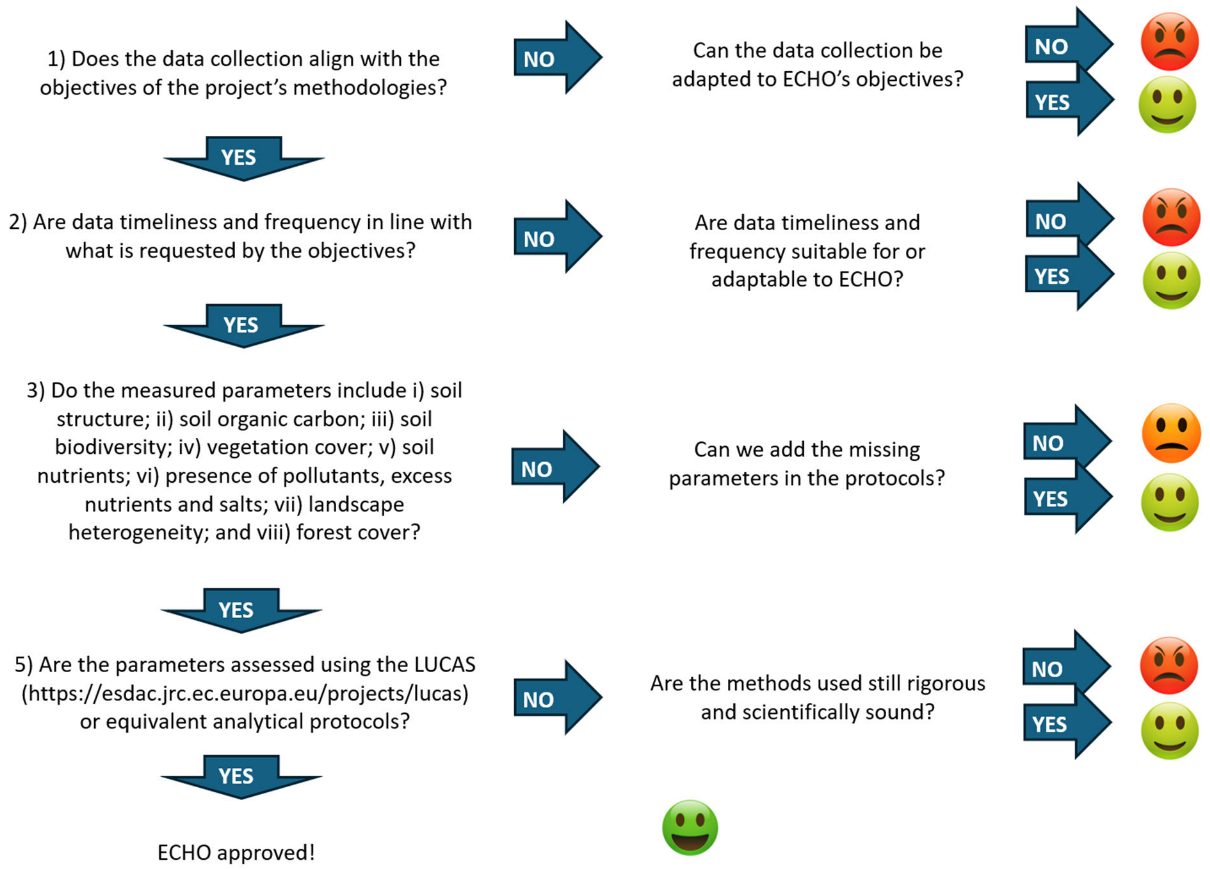

Fig. 4. Assessment flowchart for the evaluation of the dimension "Relevance".

5.2. Accuracy

Two crucial, task-independent measures of data quality—accuracy and bias—often prompt scepticism among scientists, making them critical in evaluating the potential of CGD in science. Accuracy refers to how closely a measurement matches its true or expected value. Higher precision and reproducibility of results do not necessarily indicate higher accuracy but can contribute to it. Accuracy can be assessed by comparing reference values to actual values.

Bias represents systematic errors in a dataset that consistently skew the data in a particular direction. High-quality metrics for data quality involve comparing CGD with professionally produced data and distinguishing between variability attributable to different observers and variability attributable to observer experience.

Evidence indicates that volunteers can produce high-quality data, especially in simpler tasks and when volunteers are experienced. Techniques to enhance data accuracy and address bias include iterative project development, volunteer training and testing, expert validation, replication across volunteers, and statistical modelling of errors. Therefore, each CS dataset should be evaluated individually based on its project design and application to determine its quality and reliability (Swanson et al., 2016) (Fig. 5)

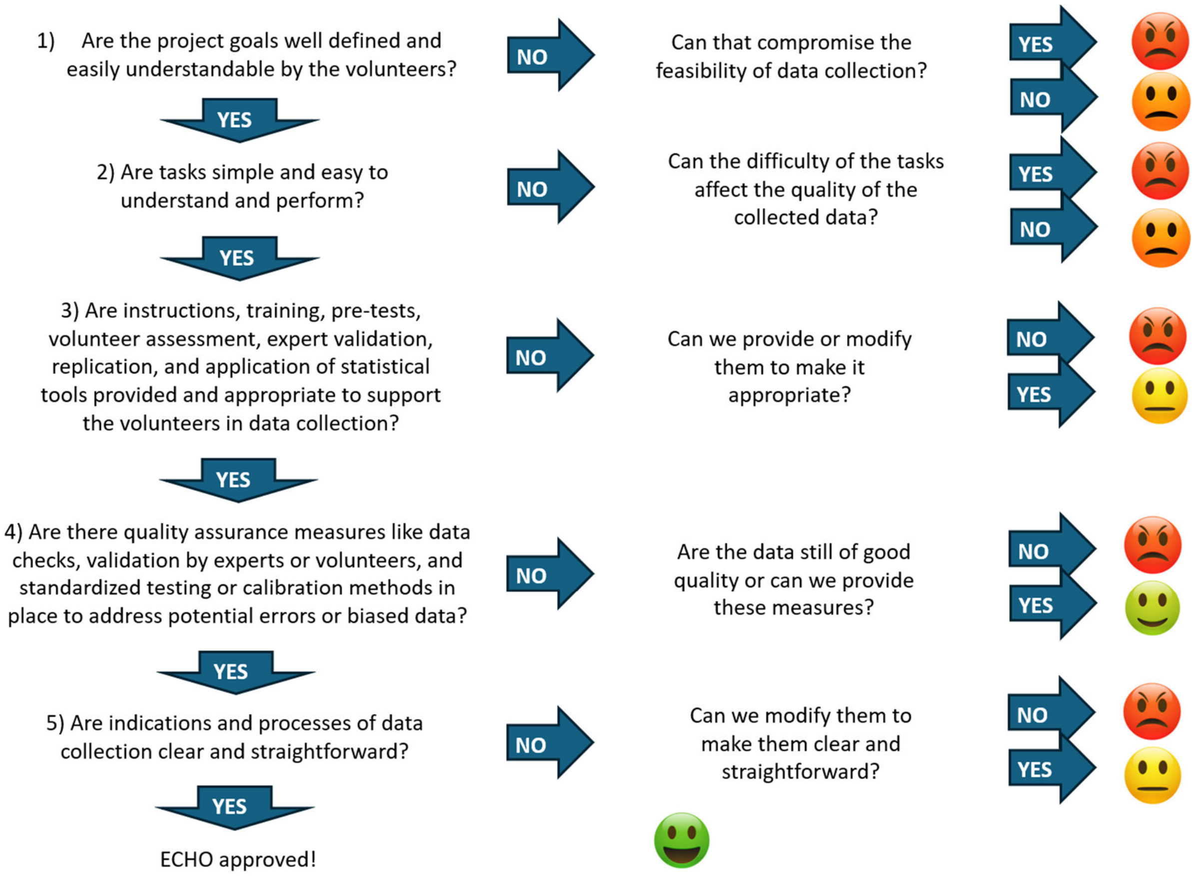

Fig. 5. Assessment flowchart for the evaluation of the dimension “Accuracy”.

5.3. Accessibility

Accessibility refers to the ease of obtaining data. Measuring this dimension might include evaluating dissemination formats, metadata availability and user support services. It's important to note that not all CGD might be publicly and readily available. The use of CGD should convene dialogues and data-sharing sessions with CGD providers who are willing to share their data, to properly assess the accessibility of the data, while also respecting privacy terms of CGD producers (Fig. 6).

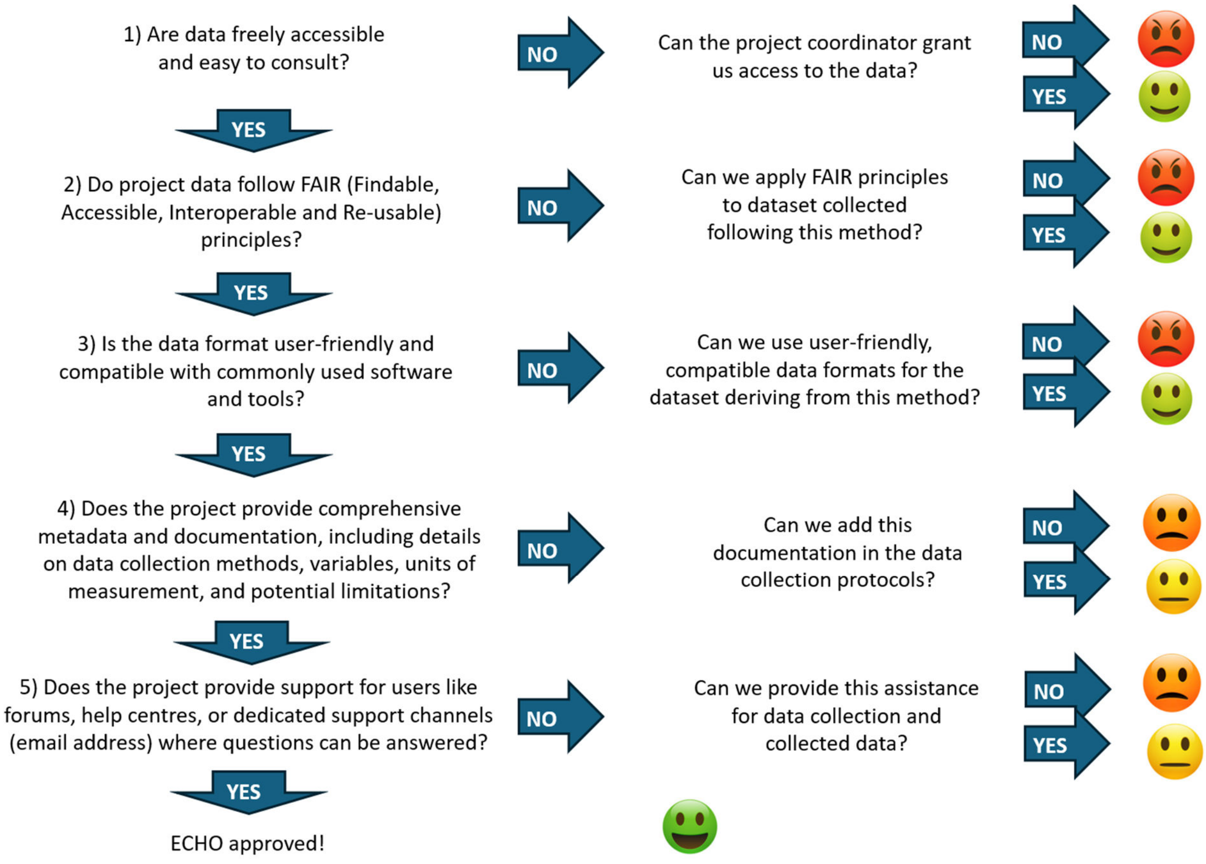

Fig. 6. Assessment flowchart for the evaluation of the dimension "Accessibility".

5.4. Coherence

Coherence reflects the degree to which data are logically consistent and can be combined with information from different sources or time periods. It can have four subdivisions: within a dataset, across datasets, over time and across different geographical units (such as municipalities, states, countries, etc). Assessing the scientific coherence of CGD is a multifaceted process that requires a combination of structured frameworks, clear communication, community engagement, and ongoing monitoring. It is essential to find a balance between empowering citizen scientists and maintaining data quality and coherence (Fig. 7).

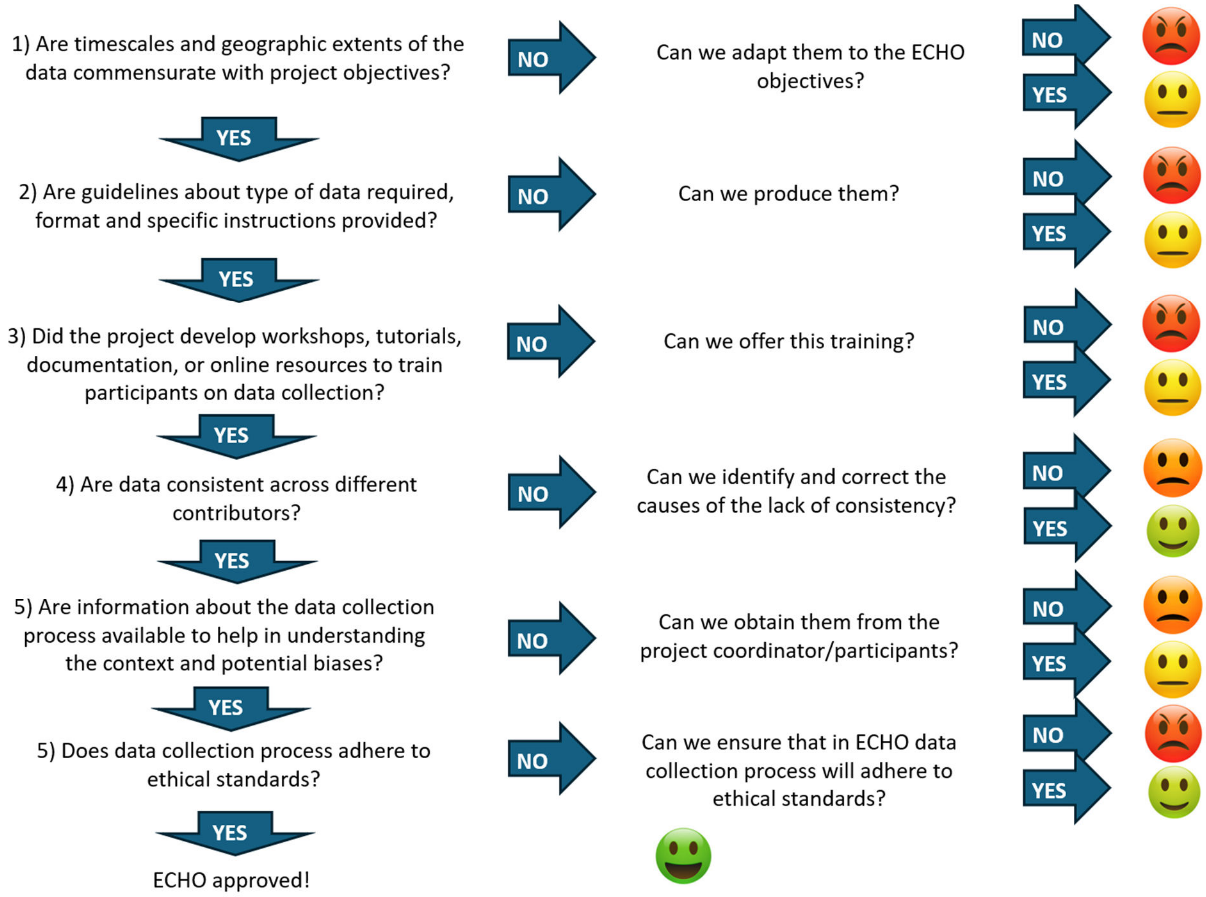

Fig. 7. Assessment flowchart for the evaluation of the dimension "Coherence".

5.5. Interpretability

Interpretability indicates the ease with which data users can understand and correctly use the data. The degree of interpretability is primarily determined by the clarity of data definitions, the target audience, the terminology underlying the data and information describing the limitations of the data. Assessing the interpretability of CGD science involves evaluating how well the results and insights derived from the data can be understood and explained (Fig. 8).

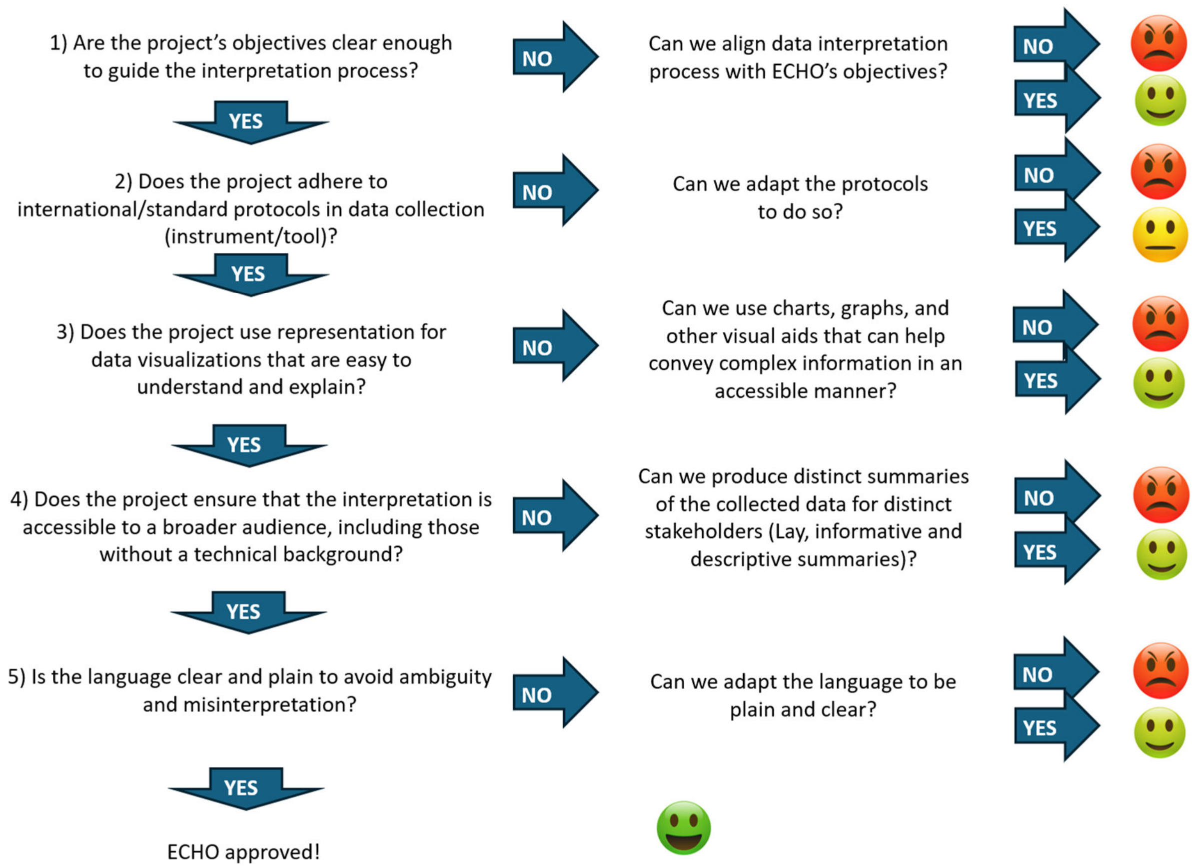

Fig. 8. Assessment flowchart for the evaluation of the dimension "Interpretability".

5.6. Compatibility

Compatibility (also known as inter-operability) evaluates whether different systems, applications, or components can work together effectively without issues or conflicts (Fig. 9). LUCAS is a European initiative, led by the European Commission's Joint Research Centre (JRC), that collects harmonized data on land cover and land use across European countries. Data collected by LUCAS is often available in vector formats, typically in Geographic Information System (GIS) file formats. While the specific data format used by LUCAS may vary, commonly used vector formats include:

1. Shapefile (.shp): A widely used vector format in GIS that consists of multiple files (.shp, .shx, .dbf, etc.) to represent geographic features.
2. GeoJSON (.geojson): A lightweight and human-readable format for encoding geographic data structures.
3. KML (Keyhole Markup Language) (.kml): An XML-based format often used for displaying geographic information in applications like Google Earth.
4. GML (Geography Markup Language) (.gml): An XML-based format for encoding geographic information, including both geometry and attributes.

5. Spatial Database Formats (e.g., PostgreSQL/PostGIS, MySQL): LUCAS data may also be stored in spatial databases, enabling efficient querying and analysis of geographic data.

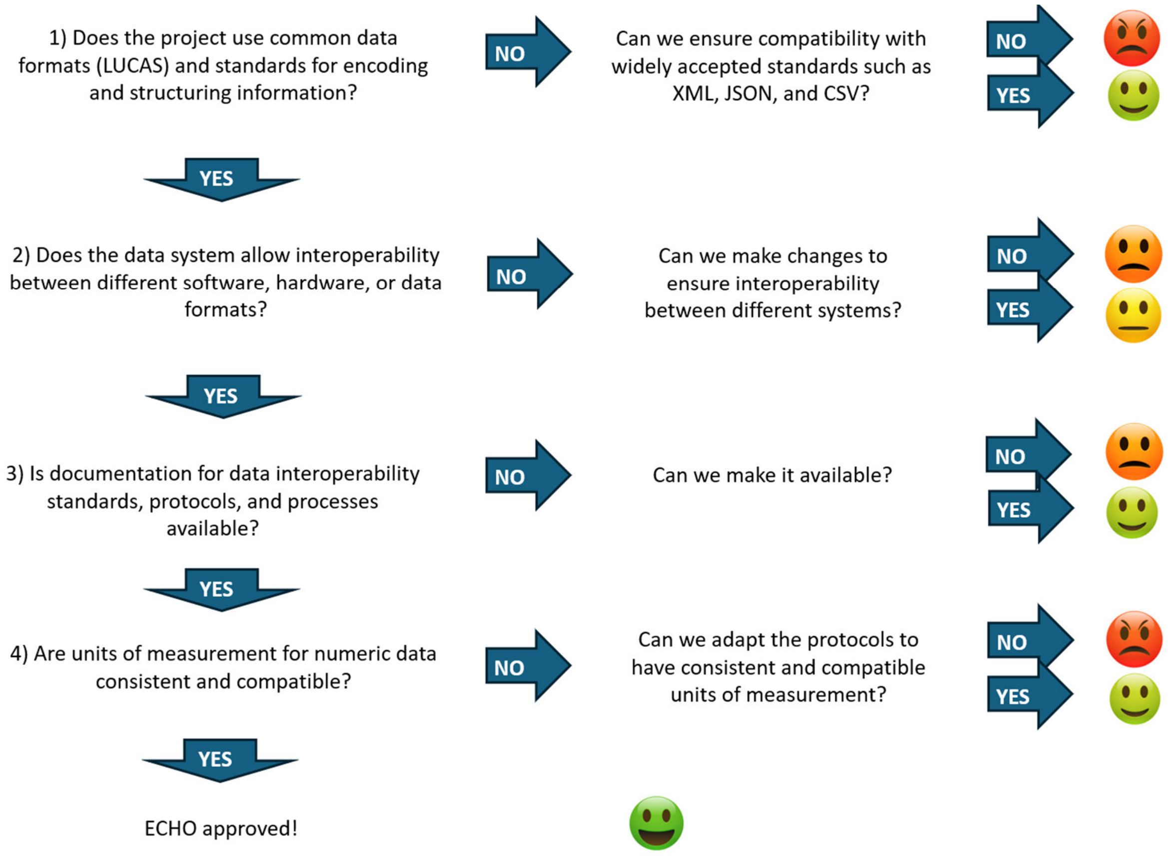

Fig. 9. Assessment flowchart for the evaluation of the dimension “Compatibility”.

5.7. Adaptation to the biogeographical region and soil types

Adapting soil indicators data to biogeographical regions and soil types is a crucial step for understanding the diversity of soil characteristics across different geographic areas. This adaptation enhances the applicability and relevance of soil data for specific ecosystems and supports informed decision-making in sectors like agriculture, land management, and environmental conservation. Achieving this adaptation involves a multidimensional approach, integrating various data sources, employing advanced analytical techniques, and incorporating local knowledge. This comprehensive adaptation enhances the usability and relevance of soil data in diverse ecosystems and facilitates informed decision-making in land management and environmental conservation endeavours (Fig. 10).

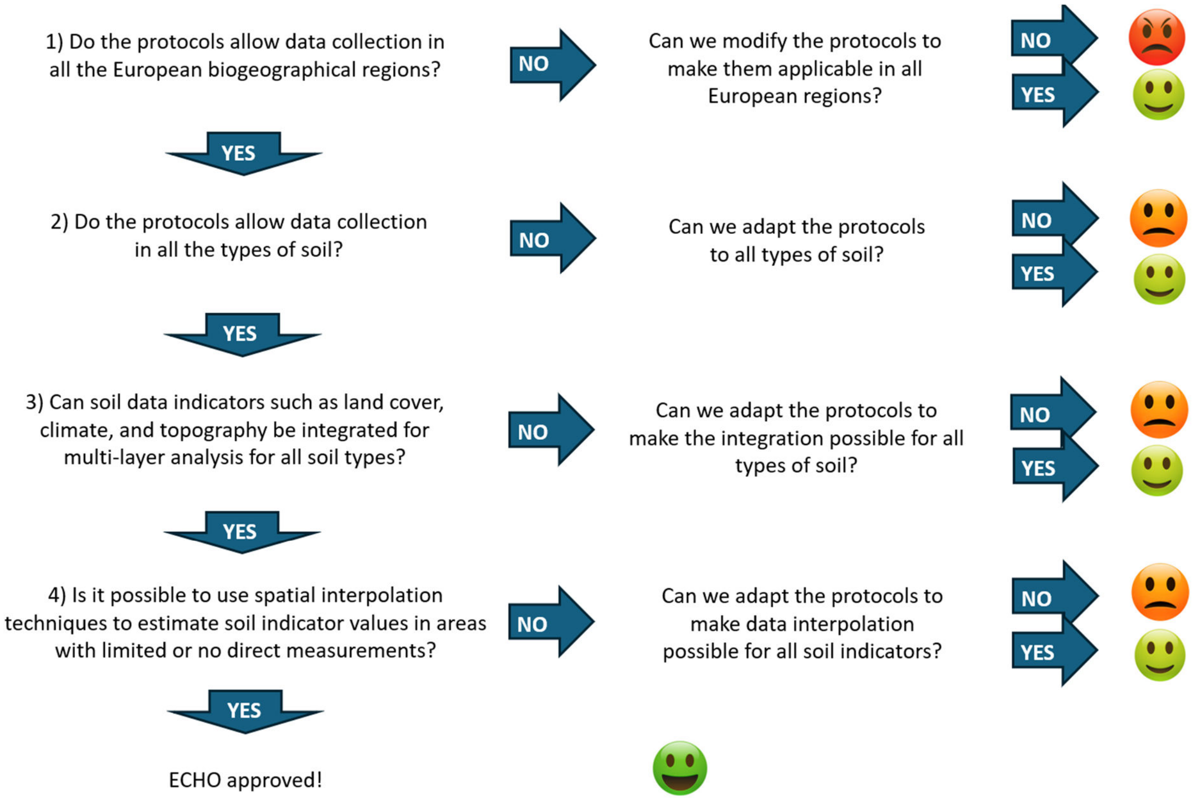

Fig. 10. Assessment flowchart for the evaluation of the dimension "Adaptation to the biogeographical region and soil types".

6. Implementation of the assessment framework

The data CGD quality assessment framework can be tailored according to the different contexts and goals of CS projects and is aligned with the Ten Principles of Citizen Science developed by the ECSA (2015) which are recognized key principles that the community considers as pillars of good practice in CS. This open framework will be applied to the previously identified initiatives outlined in Task 2.1 "Selection of citizen science methods for monitoring soils". This application will involve a detailed evaluation of the data quality from these initiatives using the established criteria in the framework. Based on the initial results generated from this assessment, the framework will be refined and adjusted to better suit the specific characteristics and challenges identified in the data. This iterative process ensures that the assessment framework remains robust, comprehensive, and aims to effectively measure and enhance the quality of CGD across different initiatives.

7. References

Balázs, B., Mooney, P., Nováková, E., Bastin, L., Jokar Arsanjani, J. (2021). Data Quality in Citizen Science. In: Vohland, K., et al. The Science of Citizen Science. Springer, Cham. https://doi.org/10.1007/978-3-030-58278-4_8

ECSA (European Citizen Science Association) (2015). Ten Principles of Citizen Science. Berlin. http://doi.org/10.17605/OSF.IO/XPR2N

Haklay, M., Dörler, D., Heigl, F., Manzoni, M., Hecker, S., Vohland, K. (2021). What Is Citizen Science? The Challenges of Definition. In: Vohland, K., et al. The Science of Citizen Science. Springer, Cham. https://doi.org/10.1007/978-3-030-58278-4_2

Ibercivis Foundation (2023). Preliminar table of Citizen Science initiatives for monitoring soil health (Preliminar Version) [Data set]. Zenodo. DOI: 10.5281/zenodo.10218825

Swanson, A., Kosmala, M., Lintott, C., Packer, C. (2016). A generalized approach for producing, quantifying, and validating citizen science data from wildlife images. Conservation Biology 30(3), 520-531. https://doi.org/10.1111/cobi.12695

Wiggins, A., Newman, G., Stevenson, R.D., Crowston, K. (2011). Mechanisms for Data Quality and Validation in Citizen Science, 2011 IEEE Seventh International Conference on e-Science Workshops, Stockholm, Sweden, pp. 14-19. doi: 10.1109/eScienceW.2011.27.