G-reqs, a New Model Proposal for Capturing and Managing In Situ Data Requirements: First Results in the Context of the Group on Earth Observations

Joan Maso [1,*], Alba Brobia [1], Marie-Francoise Voidrot [2], Alaitz Zabala [3] and Ivette Serral [1]

1   Grumets Research Group, CREAF, Edifici C, Universitat Autònoma de Barcelona, 08193 Cerdanyola del Vallès, Spain
2   Open Geospatial Consortium Europe, Technologielaan 3, 3001 Leuven, Belgium
3   Grumets Research Group, Geography Department, Universitat Autònoma de Barcelona, 08193 Cerdanyola del Vallès, Spain
*   Correspondence: joan.maso@uab.cat; Tel.: +34-935811771

Abstract: In the field of Earth observation, the importance of in situ data was recognized by the Group on Earth Observations (GEO) in the Canberra Declaration in 2019. The GEO community focuses on three global priority engagement areas: the United Nations 2030 Agenda for Sustainable Development, the Paris Agreement, and the Sendai Framework for Disaster Risk Reduction. While efforts have been made by GEO to open and disseminate in situ data, GEO did not have a general way to capture in situ data user requirements and drive the data provider efforts to meet the goals of its three global priorities. We present a requirements data model that first formalizes the collection of user requirements motivated by user-driven needs. Then, the user requirements can be grouped by essential variable and an analysis can derive product requirements and parameters for new or existing products. The work was inspired by thematic initiatives, such as OSCAR, from WMO, OSAAP (formerly COURL and NOSA) from NOAA, and the Copernicus In Situ Component Information System. The presented solution focuses on requirements for all applications of Earth observation in situ data. We present initial developments and testing of the data model and discuss the steps that GEO should take to implement a requirements database that is connected to actual data in the GEOSS platform and propose some recommendations on how to articulate it.

Keywords: in situ; requirements; datasets; Earth observations; Group on Earth Observations

1. Introduction

To observe the Earth, several techniques are used, including modern drones and remote sensing satellites. However, nothing can substitute for the observations done in situ on the ground. Each type of data has pros and cons, but together they can complement and leverage the value of each other. Due to their nature, these in situ observations are recorded by many actors ranging from professional scientists and citizen scientists, but their value goes beyond their original collection purpose. Often, observation campaigns are executed for limited space and time and lack the necessary financial support for continuous and global activity. However, if there was a way to capture different potential usages and requirements for in situ data collections, some of them could be identified as essential and its continuous monitoring and sustainable management justified.

The process of conversion of data to actionable knowledge in a distributed data services environment can be described as a sequence of steps that starts by defining the scientific question (and the problem to solve), followed by the formulation of a hypothesis, the development of a model or an analysis procedure, and finally the definition of the data requirements [1]. Only after knowing the data requirements will data users be able to search for the data they need among data providers' catalogues. In this case, data requirements are

specific to a user application and the parameters of the requirements are used to formulate a query to a data catalogue. Then, data returned by the catalogue will be used as input for an analysis and will help to demonstrate or invalidate the hypothesis.

Concrete data requirements are provided in the context of a need. There can be many kinds of needs, such as 3D modeling of the built environment [2]. Quite often, the need can be elaborated as a sequence of tasks that should be executed to solve a problem. Tasks support the users' cognitive process, facilitate the expression of parameterized user requirements, and capture the problem-solving knowledge of users. Tasks can be technically implemented by using services or workflows that will be executed to complete them [3]. A formal process of describing tasks based on data requirements has to be built on top of a set of concepts (ontology). These parameters defining data requirements should be usable in searching for the geospatial data inputs [4]. Neither geospatial metadata (describing data sources) nor the common discovery methods (in current geoportals) are organized and processed with a task-specific view. Without task-specific contexts, the gap between a task and the geospatial data becomes an obstacle that hinders quick discovery and access to geospatial resources. Instead, geospatial data resources are described by metadata and discovered through a keyword-based search function and spatiotemporal filters exposed by geoportals. Most of the time, these keywords are related to the measurement or production process rather than to the needs or tasks the data can support. For example, when undertaking a specific task about "habitat preservation at the national level", we cannot directly submit this expression in a geospatial data catalogue. Instead, we commonly submit a query clause specifying data requirements such as "spatial resolution ≤ 100 m" and "theme = habitat descriptions" or "theme = land cover classes". Thus, users have to first translate the task at hand into geospatial data characteristics and then translate these characteristics into a concrete query [5]. The translation is not direct because it depends on spatiotemporal factors. For instance, using remote sensing images for the "land cover classification" task at the local scale or at a global scale will require different spatial resolutions.

It is fundamental to associate every requirement of the measured variable. One possible dictionary of data themes is provided by the essential variables (EVs) framework. EVs are the minimal set of variables that determine a system state. They are crucial for predicting the system developments and allow calculating metrics that measure the trajectory of the system [6]. EVs aim to adequately describe the various subsystems that constitute the Earth system (e.g., atmosphere, biosphere, geosphere, and hydrosphere) and to ensure that all potential users have a minimum set of observational data that they may need [7]. The concept of EVs has been adopted by some communities. The Group on Earth Observations (GEO) is an association of members and participating organizations that since 2005 has worked in a coordinated manner to increase the accessibility, use, and openness of data to extract maximum benefits from Earth observation (EO) data (both satellite and in situ) and to impact policy by enabling a data-driven decision-making process. As a GEO activity, the Group on Earth Observations Biodiversity Observation Network (GEOBON) community defined a set of EVs called essential biodiversity variables (EBVs), and they are now identifying requirements for each of them and looking for the products and matrices to describe them. In particular, Ref. [8] discusses the ideal and minimum requirements for the species' distribution and abundance at a global scale. The Global Climate Observing System (GCOS) community initiated the process of essential climate variables (ECVs) years before, and has released a list of products describing its requirements [9]. The GEO essential variables community activity provides a forum in which to discuss the completion of the essential variable (EV) framework by adding other domains to exchange knowledge, experiences, and methodologies in EVs' definition, as well as to analyze the usefulness of some EVs in monitoring the sustainable development goals (SDGs).

The World Meteorological Organization (WMO) and cosponsored programs have developed the Observing Systems Capability Analysis and Review (OSCAR) tool as an official repository of requirements for each ECV. OSCAR is the foundation of the rolling requirements review (RRR) process, overviewed by the Inter-Program Expert Team on

Observing System Design and Evolution [10], in which requirements are regularly reviewed by groups of experts. Sometimes, the data product that matches the requirements already exists, but in some cases a need to create a specific data product emerges. For example, Ref. [11] detected that certain variables measured from satellites (such as water vapor) generate products with quality parameters that do not fulfill the OSCAR requirements for the considered application. The paper discusses how to solve this gap by using a model that can deduce water vapor from accumulated rain. The OSCAR system is an evolution of a requirements database called CORL, developed by NOAA in the context of NOSA [12]. The system evolved into the NOAA Consolidated Observation User Requirements List (COURL) [13]. This development is still active as part of the OSAAP [14]. Requirements Capabilities and Analysis for Earth Observations (RCA-EO) is a U.S. Geological Survey (USGS) National Land Imaging (NLI) program partnership with other federal agencies used to document user requirements for Earth observation data (https://www.usgs.gov/rca-eo; accessed on 12 February 2023). The Copernicus In Situ Component Information System (CIS^2) [15] is an evolution of the COURL and OSCAR model adapted to the needs of the Copernicus services. The Copernicus services provide a myriad of remote sensing-derived products and numerical models that are calibrated and validated by using in situ data. For each product, the CIS^2 database collects in situ data requirements which are expressed by the Copernicus-entrusted entities/data providers and parameterized in terms of accessibility, usability, and quality criteria. The catalogue can be explored here: https://cis2.eea.europa.eu/about (accessed on 10 February 2023).

With specific regard to data quality aspects, requirements are commonly expressed with parameters for positional, temporal, and attribute accuracy [16]. However, user requirements may consider parameters that characterize the data that are not commonly considered as data quality but are critical for a transversal usability of the product, such as horizontal resolution, vertical resolution, observing cycle, or timeliness.

Challenges related to in situ data collection were recognized in the GEO Canberra Declaration [17] as well as the need to manage end-users' requirements to better serve the community. In response, the GEO in situ subgroup was created to strategically work to identify the main barriers in data sharing and management and to coordinate existing networks of in situ data providers in specific domains at different scales. Indeed, GEO's engagement priorities include support for the UN 2030 Agenda for the SDGs [18] across eight topics or societal benefit areas (SBAs) where Earth observations play a key role. To measure the progress towards the SDGs' targets, the SDGs' policy indicators are defined. The computation of SDGs' indicators generates information needs that result in workflows that require data. The United Nations maintains one metadata document per indicator, detailing the data needs and enumerating the procedure by which to elaborate on the indicators from existing or nonexisting data. They mainly use national statistical agencies' data but the usefulness of EO data in general, and of in situ data in particular, was also identified in a number of cases [19]. GEO can contribute to collect and serve these needs through its initiatives such as the GEO Global Agricultural Monitoring (GEOGLAM), or the previously mentioned GEOBON, as these communities are developing their own thematic in situ databases and can help define data requirements and serve sustainable development goal (SDG) indicator implementations.

This paper describes Geospatial requirements (G-reqs), a database of in situ data requirements. The challenge is to enable the reuse of data beyond usages that the data provider was not able to consider. A novelty of the approach empowers the user to provide data requirements, based on the concept of needs that can be broken into tasks that require data. In contrast with most of the existing systems, where requirements are collected and managed by a data provider, G-reqs is contributed as a tool for the GEO in situ subgroup, in which both data providers and data users are represented. In a second phase, it enables one to identify the datasets' value by quantifying the number of occurrences in which datasets are requested. Technically, the access to the database follows the new OGC APIs design pattern.

In this paper, the Materials and Methods section describes the G-reqs database structure and its implementation as a Web form. The First Results section explains the testing outcomes based on the pilots of the Horizon 2020 e-shape project. The Future Work section explains the processes and tools necessary to populate and maintain the database, proposes a process by which to analyze the collected user requirements and to extract product requirements for new and existing products, and formulates a set of recommendations to implement the G-reqs in the GEO. The paper ends with the Conclusions section.

2. Materials and Methods

In the in situ subgroup of GEO discussions, the need for an in situ requirements database was detected as an instrument with which to justify and structure the efforts in organizing the availability of in situ data. The European Environmental Agency (EEA) already designed and continues to maintain the CIS^2 requirements databases for the Copernicus in situ component, which targets the remote sensing calibration and validation needs for the high-level products offered in the other Copernicus services. The EEA proposed to extend the CIS^2 into a new system that covers the global needs of GEO. This section describes the resulting new data model and its initial implementation as a Web form, called G-reqs.

From the analysis of the existing data requirement databases, we found they mainly follow a producer-centric approach. For example, when the weather community agrees on the parameters describing a set of required datasets and stores it in the OSCAR system, they define a set of products that WMO should make available. In contrast, the CIS^2 starts by the Copernicus services needs that are a concrete community of users of the in situ data. In this paper, an in situ requirement system called G-reqs that focuses on user requirements for a wide range of GEO users is presented.

G-reqs was designed by using unified model language (UML) class diagrams. UML is a standard visual modeling language intended to be used for the analysis, design, and implementation of software-based systems [20]. It defines several types of diagrams, including the class diagrams used here. Classes are represented as boxes with a list of properties inside. In this case, a class will be implemented as a table in the requirements database. A class is related to other classes by aggregation (lines starting with a diamond and ending with an arrow cap). In our case, aggregations represent relations in a relational database. In the UML diagrams, we want to emphasize that some classes are copied and modified from the CIS^2 requirements data model. In these cases, in the G-reqs data model, a general class with some properties from the CIS^2 data model is extended by adding more properties in a G-reqs specialized class. In UML, a class is a specialization of a more general class when there is a relation represented by a line ending in a triangular arrow cap pointing to the general class.

The user approach described in G-reqs is based on three main objects: *Need*, *Task* and *UserRequirement* (see Figure 1). The user (represented by a "GEOSS user" in the model) has a concrete applied *Need* (problem or issue that can be addressed by using in situ data). There are many types of *Needs* including calibration and validation of remote sensing products (*Remote Sensing CalVal*), calibration or validation of a modeling software (*Model Input CalVal*), demonstration of a scientific hypothesis (*Scientific Research*), aggregating it in a more general data service (*Enrich Data Service*), provide a commercial service derived from the data (*Commercial Prod Serv*), calculate a policy monitoring indicator (*Policy Indicator*), or assist in a decision-making process (*Decision Making*). To cover a need, a set of *Tasks* are executed, some of them requiring in situ datasets describing the status of a particular *Essential Variable*. Each in situ dataset is described as a *UserRequirement* in the model. The properties of a requirement are organized into the three main components of geospatial information: spatial, thematic, and temporal. A user requirement defines spatial extent (*area code*) and temporal extent. Uncertainties for the three components are described, as well as the updated frequency and the timeliness (defined as the time lapse between the data measurement and the data accessibility). In addition to parameters that

quantitatively describe the needed data, the measurement strategy, the license, and the needs to access the data are also captured.
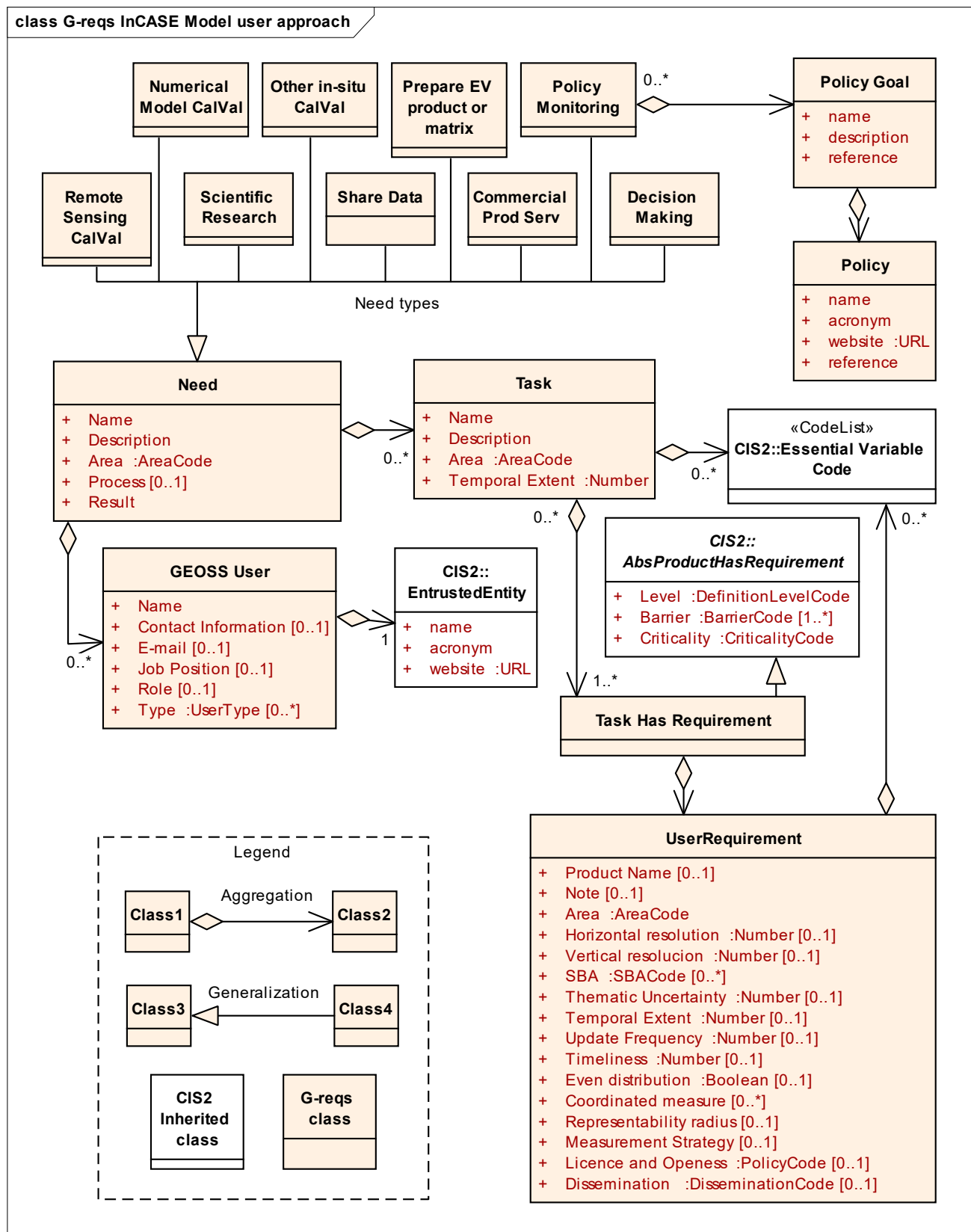
Figure 1. UML Class diagram of the user requirements data model.

Three properties that are specific to in situ data are introduced: even distribution, coordinated measures, and representability radius. For in situ data, an even distribution of observations is not always required, in particular, if the data comes from citizen science, as it is mostly collected around citizens' homes or in focused campaigns. However, statistically sound studies may require even distribution of measures, such as the variability of cropland topsoil properties [21]. Sometimes, when detailed observations are difficult or expensive to obtain, it may be better to focus on particular stations that are considered representative or a reference for long series. In this case, it is preferable that observations of different themes are recorded in a coordinated or holistic manner, and in a way that localized comprehensive studies can be conducted and their results extrapolated to the rest of the territory. The critical zone observatories represent an example of this spatially focused practice [22,23]. As a final aspect, due to its nature, in situ measures are done by observing a single location (e.g., a point or transect). However, these locations can be selected in a way that makes the measures representative over a radius around that location. For example, in phenological studies that combine in situ observations and remote sensing, it has been suggested that an in situ observation should be done in a patch of homogeneous vegetation that covers at least a 3×3 pixel area in the remote sensing counterpart product [24].

Once the properties of the requirements for a dataset are completely defined, a dataset fulfilling completely or partially the requirements may or may not be found as expressed by the existence of a *Data Meet Requirement* class that connects to existing in situ datasets, as explained in the following sections of this paper.

G-reqs proposes to capture concrete user requirements via a simple tool that supports users' needs' description and its translation into comprehensive data requirements. The tool formulates a set of questions organized in a Web-accessible form that the user should answer. Each question is accompanied by an exemplary response. The tool is an implementation of the G-reqs user requirements model targeting EO data users. The tool access URL (https://g-reqs.grumets.cat; accessed on 10 March 2023) was distributed among the GEO community of users, starting with the regional EuroGEO through the European Union e-shape project. The form introduces the conceptual model approach to help users understand the goal and the process and eventually identify requirements and report on them. Semiopen questions allow users to describe one specific *Need* and one or more concrete *In situ requirements*. A first set of questions asks to describe the *Need* by providing answers to the questions in Table 1. In order to make the form short and easy to populate, the enumeration *Tasks* that should be executed to cover the *Need* was deliberately not included in the form; instead, there is a question about the "process" used to realize the need that addresses the tasks sequence. By thinking about how to cover the need, we assume that the user already formulated the tasks and he/she is ready to directly report on in situ data requirements derived from the tasks. The set of questions describing a concrete in situ data requirement (that in turn technically describes the properties of a potential in situ dataset) are listed in Table 2. For each *Need*, a maximum of five *In situ requirements* can be provided.

Table 1. Questions about the *Need*.

Question	Example
Short name.	Water quantity in Greece.
Short description.	Water flow and rainfall in Greece.
Process or method used to realize the need.	Hydrographic model that requires water flow and rainfall as input data to assess water quantity.
Geographical area of scope.	Greece.
Result.	Hydrographic numerical model cal/val.

The information gathered from users allowed validating the G-reqs model and capturing reference requirements from the GEO community as a basis for a standard collection of user requirements on in-situ data.

Table 2. Questions about the *In situ requirement* [1].

Question	Example
Short name.	Water flow in Greece.
Short description.	Water flow in Greece to assess data quantity.
Essential variables, if any, related to the in situ data.	EWV runoff/streamflow/river discharge.
Type of data access expected.	Discovery service, download service, view service, COG file.
Related GEO "societal benefit area" or topic.	Water resource management.
Thematic uncertainty (thematic error) acceptable expressed in units of measure.	~4 m^3/s.
Update frequency or how often data needs to be measured.	Every 15 days.
Timeliness or delay between measurement and results accessibility is acceptable.	In 1 h.
Temporal extent or period covered by the data, if historical data is relevant.	Only current data needed.
Even distribution or spatial uniformity of samples is relevant.	Uniform distribution along the river.
Type of coordinate measurement if other in situ measurements are needed to be taken in the same place in a coordinated way, if relevant.	Meteorological stations, eddy covariance flux-station, soil characteristics.
Representability radius or minimum distance around the in situ measurement, if relevant.	A weather station is located in a flat area free of trees and with grass so, if the measurement was done at 200 m distance, it will be equivalent.
Horizontal resolution expressed in units of measure.	Samples are taken as 15 km average distance.
Vertical resolution, if relevant.	Every 200 m.
Reference of existing in situ dataset that partially fulfill the requirement if available and aspects missing in the dataset referenced that prevents using it.	https://www.riob.org/IMG/pdf/RBMP_Greece_April_2013.pdf. Not current.

[1] These questions are repeated for each requirement with a maximum of five requirements.

3. First Results

The G-reqs in situ requirements model was tested in the Horizon 2020 e-shape project where showcases in agriculture, health, energy, ecosystems, water, disasters, and climate are developing a total of 37 different pilots in support of the European contribution to GEO. The G-reqs Web form was sent to the pilots (approximately one third of the pilots responded) and introduced in a European GEO symposium and the first results are analyzed here.

The main needs collected were the calibration, validation, and quality control of remote sensing data (in particular for biodiversity and water quality), evaluation of numerical model outputs (in agriculture, weather, and air quality), scientific research (effects of pollution on monuments and city populations and status of water bodies), generalization of low-resolution data (pollutants' distribution), development of indicators (for climate change), prepare EVs products (EOV and EBV) and decision-making (in renewable energy in conjunction with satellite data). The main variables required to collect are air quality and pollution (NO_2, CO, SO_2, O_3 PM2.5, PM10), water quantity and quality (chlorophyll-a, suspended matter, turbidity, phycocyanin, cyanobacteria, nutrients, dissolved oxygen, DOC, presence of scum, water temperature, temperature profiles, pCO_2, currents, and salinity), biodiversity (tree species in cities, effects of drought in trees, biomass, phenology), oceans (wind speed and SST), agriculture (soil, harvest and yield, crop height, number of leaves, vitality), energy (downwelling solar radiation at surface, total solar irradiance, cloud properties, albedo) and weather and climate change common measurements. Talking about the concrete in situ requirements provided and their relationship with the SDGs (see Figure 2a), the eLTER network data requirements address SDG2, SDG3, SDG6, SDG7, SDG11, SDG13, SDG14, and SDG15. SDG2 is addressed by crop yield data requirements. SDG11 is addressed by in situ measurements of annual mean levels of fine particulate matter (e.g., PM2.5 and PM10) in cities and has implications for SDG3. SDG13 is addressed by data requirements for analyzing alien/native species, invasive/alien species, and toxic/alien species in urban cities. With regard to the GEO Societal Benefit Areas (https://www.earthobservations.org/geo_wwd.php; accessed on 15 February 2023), the collected in situ requirements are related to biodiversity and ecosystem sustainability, public health surveillance, water resources management, sustainable urban development, and food security and sustainable agriculture (see Figure 2b).

The EVs framework is relatively new and, in some spheres (such as oceans and biodiversity), the definition of EVs was elaborated based on what was needed and not on what it measured. This resulted in some EVs that have never been produced in an operational way. As previously stated and exemplified by [8], the creation and maintenance of datasets measuring each EV is needed, and as another example in the e-shape project, two pilots reported using in situ data for creating a dataset that describes an EV. Actually, this type of need was not initially foreseen in the model and has been added to the G-reqs

after the test phase. Even if it is not their main objective, more than half of the collected requirements are connected to some sort of decision-making process. The geographical area varies between global, European, regional, national, subnational, and local depending on the type of study. With regard to temporal resolution, depending on the application, an annual value could suffice (e.g., crop models) but for others, hourly frequency is needed (e.g., air quality monitoring).
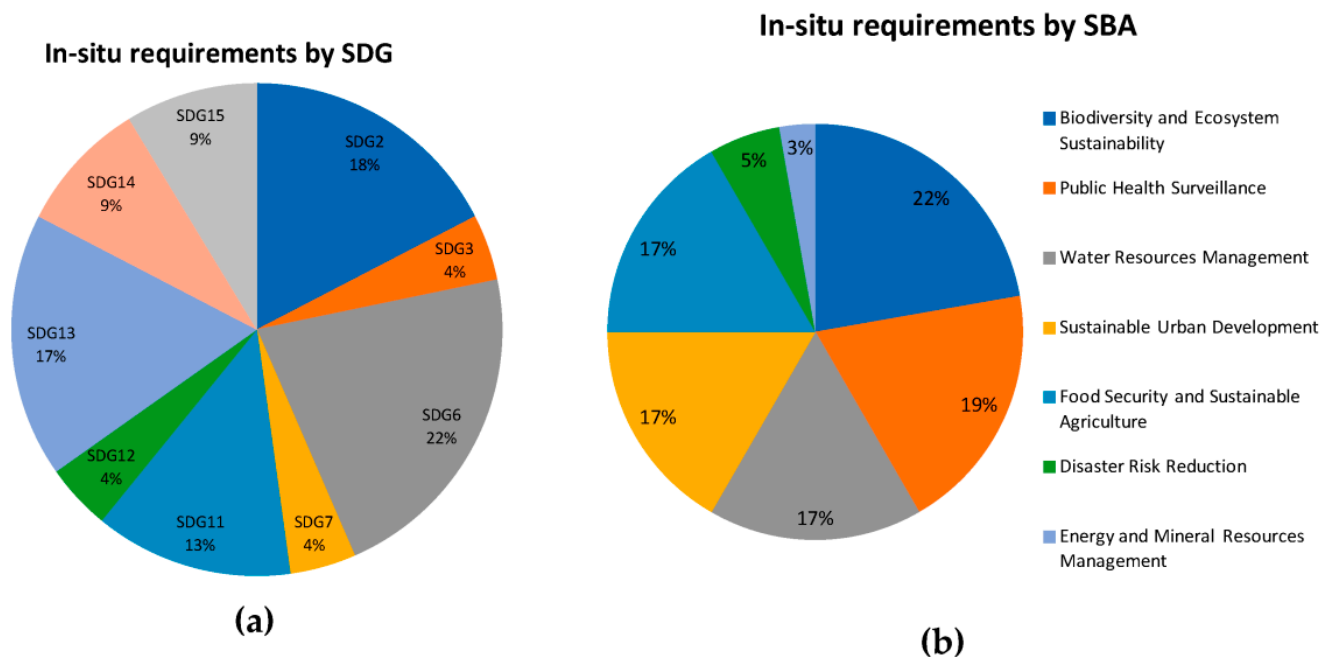
Figure 2. In situ requirements (**a**) by SDG and (**b**) by SBA.

In terms of data accessibility, we observe the request of all sorts of data access ranging from a classical HTTP download to get a CSV or a NetCDF, to a more elaborate request for the extraction of a time series of a single location or the access to a cloud service.

With regard to in situ specific properties, an even distribution of measures (responded by 80%) is needed in 50% of the cases for air quality and water quality, and coordinate measurements are required in 40% of the cases, e.g., with weather stations, soil type, and soil moisture.

The main requirements can be summarized as the accessibility and discoverability of data for air quality, a lack of easy availability of water quality and quantity (requiring contacting the data provider), spatial gaps in tree inventories resulting in cities partially mapped, insufficient resolution for physical water parameters, and too many uncertainties in air quality data as well as a lack of samples to get conclusions for losses and damages on the forest.

As part of the validation efforts, the web form included an additional question to provide feedback about the in situ data user requirements data model. Some comments suggested new need types that have been added to the model (such as the elaboration of EVs products or validation of other in situ and citizen science data) keeping the question open for allowing other unexpected types. Other comments requested to include a relation between the need and the data user, that has been applied. None of the participants provide negative feedback on how the requirements were modeled.

Two fields describing requirements are particularly difficult for the responder. The thematic uncertainty was only provided as a numeric value in 37% of the cases; two responders asked for the best possible uncertainty, and one asked for clear uncertainty documentation for the dataset, whatever it was. The representability radius was correctly responded to by 19% of the cases and was found to be difficult to respond to and sometimes

required extra comments (e.g., "avoid edges of the agricultural field"). In contrast, the update frequency and temporal extent were correctly reported by 100% and 93% of the cases, respectively.

The EVs framework covers 90% of the cases but some specific measurements do not fall in any of the current categories (e.g., effects of drought on trees, presence of scum in water). A total of 25% of the responders were not able to provide the correct answer, possibly due to their lack of EVs framework knowledge.

4. Future Work

The capacity of G-reqs to collect in situ user requirements was described in the previous section. To make G-reqs really useful, it is necessary that G-reqs also allows sharing of the data, analysis of the requirements, detection of gaps in in situ datasets, and help in making recommendations to data providers, thus promoting the discovery of fit-for-purpose in situ datasets (see Figure 3).
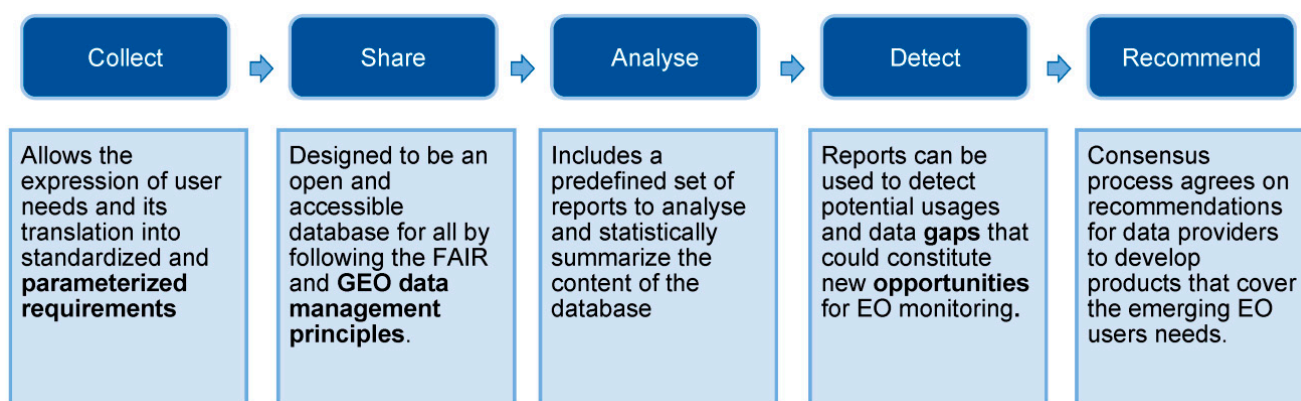
Figure 3. G-reqs functionalities and development process.

G-reqs is designed to be an open and accessible database for all. To do so, it will also adopt the GEO data sharing and data management principles [25], ensuring adequate data management of the captured requirements as well as its free access. To share the data, we offer a CSV file with all gathered requirements as a starting point, but we propose to make the requirements interoperable by means of the new OGC APIs. The proposal consists of assimilating a requirement into a geospatial feature and using a variation of the OGC API features [26] with three feature collections—one for needs, another for tasks, and another for requirements. Each one could use the item type need, task, and requirement, respectively (in the same way as the OGC API records proposes the use or item type record to implement a metadata catalogue through the OGC APIs). The response of the service could be a GeoJSON file with the geometry reflecting the polygon of the *area code* and all other attributes as properties. An extra property can be used to set links between needs, tasks, and requirements.

While the OGC API features will provide some query capabilities, G-reqs should include a predefined set of queries that result in reports to analyse and statistically summarize the content of the database. Some of the G-reqs reports can be used to group requirements by essential variable and detect new potential in situ usages and gaps in in situ observations that could constitute new opportunities for further developing Earth observation monitoring.

The G-reqs model includes the transformation of user requirements into product requirements (see Figure 4). A product requirement describing the status of a particular essential variable is agreed upon by a consensus body. The consensus body takes into consideration the user requirements that have been formulated before in relation to the EV, making the product requirement traceable to its original user requirements. While user requirement properties are expressed with a single value, most of the product requirement

properties are defined as an interval described by three values (as proposed by the OSCAR model) expressing an interval that starts at the minimum value that makes data useful for some purpose (*threshold*) and ends in the ideal value above which further improvements are not necessary (*goal*). A middle value is also provided (*breakthrough*) that is considered optimum, from a cost–benefit point of view, when planning observations. If the consensus process results in an agreement, a recommendation for producing a product is issued, and the right community or network that can potentially cover the emerging need for Earth observation is informed. Figure 5 shows the connection between the user and the producer approach based on the same essential variable dictionary and the traceability of the original user requirements from a product requirement. GEO has done a considerable effort in formalizing a set of EVs that covers most of the entire spectrum of data themes in the geosciences through the incorporation of the Essential Variables community activity in the current GEO Work Programme 2023–2025. However, while the EVs framework specifies the variables, it does not necessarily specify the requirements of the community for each product measuring EVs. G-reqs provides a tool with which to support the process of taking user requirements and formulate and prioritize product requirements for existing or future products that measure each EV in a way that covers the community needs.
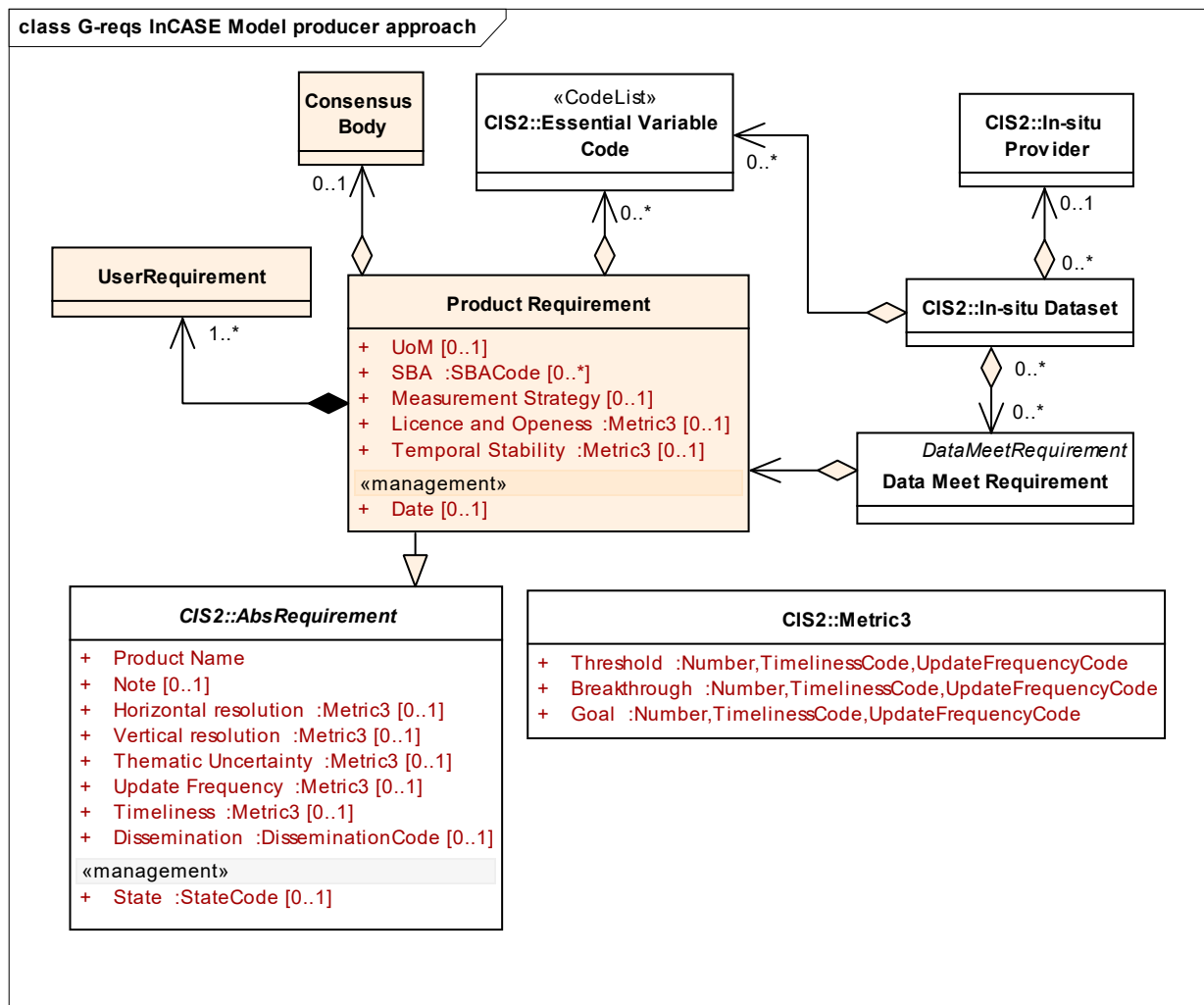
Figure 4. UML Class diagram of the product requirements data model (see legend in Figure 1).
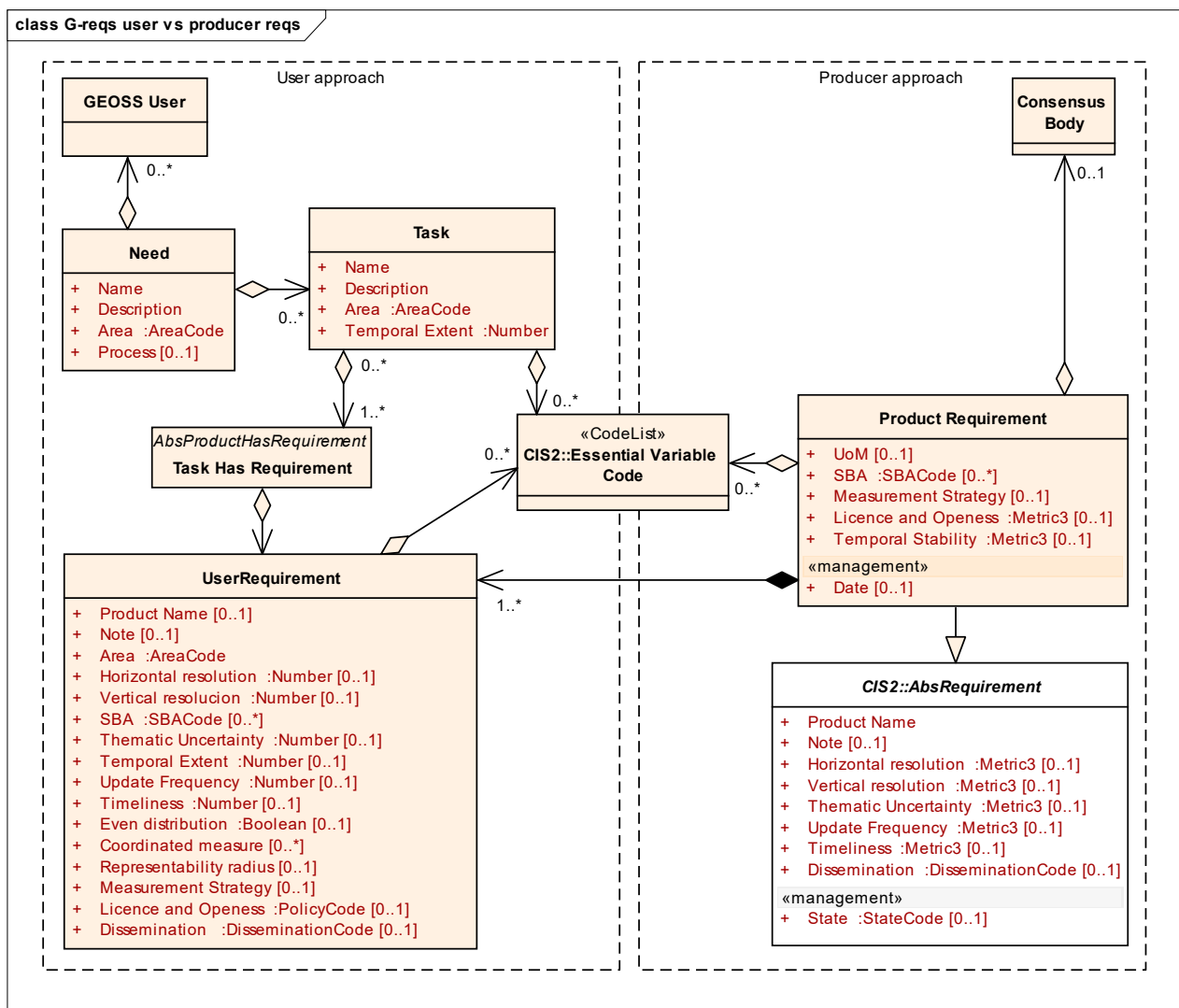
Figure 5. Connection between the user requirements and the product requirements (see legend in Figure 1).

Tentatively, the Web form includes a semiopen question about possible datasets that the user may know and that partially or completely covers their needs. Approximately half of the responses already provide a possible provider, some including a reference to a dataset. In that respect, to be able to know if a user requirement can be covered by an existing in situ dataset in GEO and to formulate a recommendation for making new in situ data accessible in GEO, we need to know about the capacity of the current GEO providers in providing in situ data. The Report on Usage of the GEOSS platform [27] indicates that 27% of the GEOSS providers integrated in the GEOSS platform are purely in situ (22% of the services), and 55% provide both types of data (57% of the services). A more detailed analysis of the number and characteristics of the in situ data providers in GEO can be extracted from an analysis of the GEOSS Yellow Pages database. Figure 6 shows the proportion of in situ data providers that declare support to an SDG and an SBA, demonstrating that all SDGs and SBAs are covered by at least one data provider, which implies a rich thematic coverage too.

While in situ data is intrinsically local, Figure 7 shows that 69% of the in situ data providers declare they maintain global in situ products. There is some chance that a concrete user requirement for in situ data could be immediately fulfilled by an already existing dataset and provided by a data provider integrated into the GEOSS platform. We have checked if the advanced search of the GEOSS portal (https://www.geoportal.org;

accessed on 5 February 2023) has the necessary options to formulate an automatic query based on the collected user requirements. Unfortunately, only a few of the characteristics are exposed as queryable parameters: in the advanced search of the GEOSS portal, it is possible to filter by spatial and temporal extent, specify a thematic area or keyword, roughly define the timeliness of the data, and determine if the data is part of the GEOSS Data CORE (a GEO concept similar to "open data"). Unfortunately, there is no way to query specifically for in situ datasets, and it is not possible to query by spatial resolution, thematic uncertainty, update frequency, or any of the three specific characteristics for in situ data: even distribution, coordinated measures, or representability radius. With the current limitations of the GEO portal advanced search, it is not possible for the G-reqs system to automatically formulate a query and determine if a dataset that fits the purpose of a user based on the reported user requirements is available. Only a partial filter per extent and timelines could be formulated, and the user should manually look for a dataset meeting the requirements among the partial query result records by carefully analyzing the description and the metadata of each individual record.
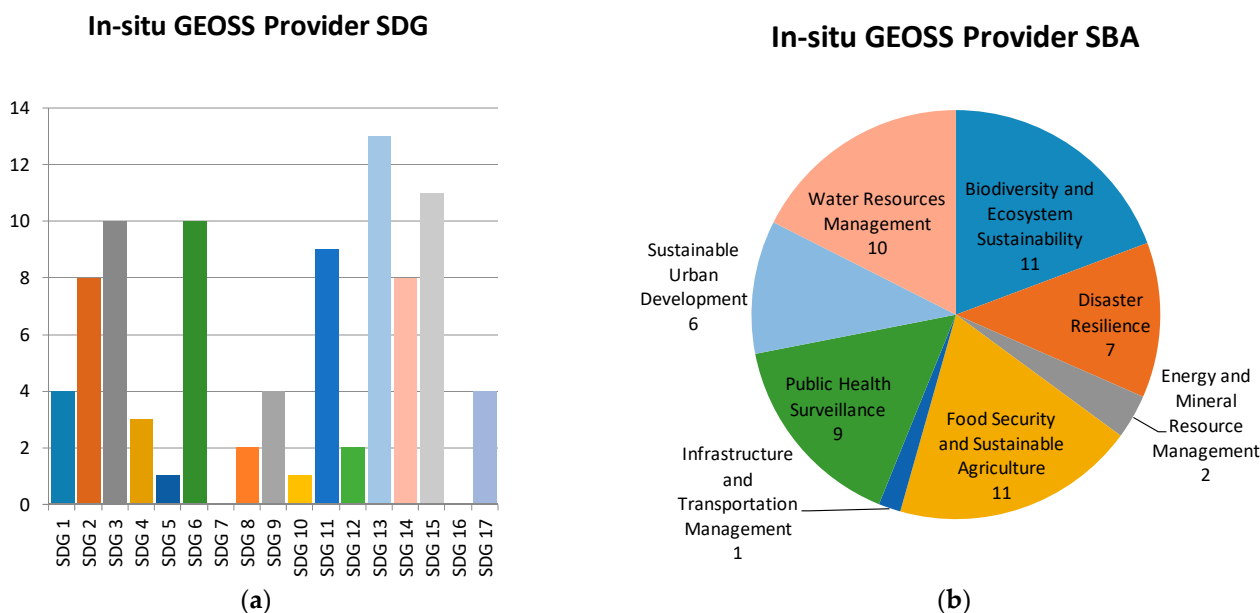
Figure 6. In situ data providers in GEOSS (**a**) classified by SDG; and (**b**) classified by SBA (data from [26]).

At the time of writing this paper, GEO had 113 members, 143 participant organizations, and 19 associates, all with appointed delegates. These delegates are a community articulated in a work programme composed of GEO flagships, initiatives, pilot initiatives (formerly known as community activities), foundational tasks, and regional GEOs. In case a user requirement is not covered by a dataset documented in the GEOSS portal, it should be possible to contact a specific member of the community to check if the data exists and assess if it could be made available or if it does not exist and should be considered in a future in situ campaign. Currently, GEO does not offer this capacity as a formal service, but the authors of this paper recommend that the capacity should be articulated in the future. Another opportunity for G-reqs consists of using the number of requirements collected on a particular data product to improve the ranking of the answers in the GEOSS portal.

This paper proposes a data model to gather geospatial data requirements based on describing needs, tasks, and requirements that are voluntarily contributed by the community. More automatic ways of populating a database could be tested, such as training a case-based reasoning (CBR) algorithm to automatically derive requirements from needs and tasks [5].

**In-situ GEOSS Provider Geographical coverage
of the organization**
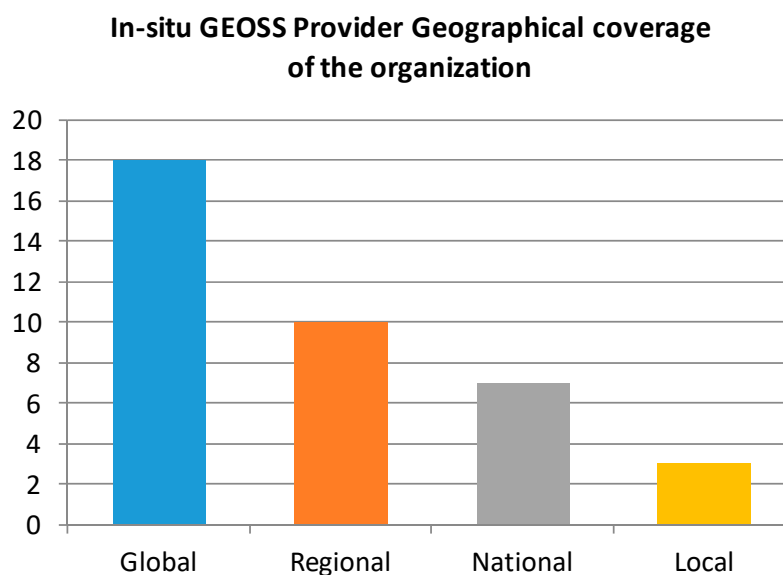
Figure 7. Data extent of some of the in situ data that providers offer in GEOSS.

5. Conclusions

G-reqs aims at facilitating the match between user needs and in situ data collections by defining a data model for user requirements and supporting the translation of user needs into user requirements via a simple Web form. The Web form acts as a user interface into the G-reqs requirements database.

From the analysis of the in situ requirements already gathered, we can conclude that users understand the model and correctly respond to questions about each parameter. The main needs collected are the calibration, validation, and quality control of remote sensing data, numerical model outputs, and generalization of low-resolution data. Even if it is not their main objective, more than half of the collected requirements are connected to some sort of decision-making process. The geographical area varies between global, European, regional, national, subnational, and local depending on the type of study. Depending on the application, an annual value could suffice, but hourly frequency is needed for other applications. Talking about specific in situ properties, uniform spatial distribution of measures is needed for air quality and water quality; coordinate measurements with weather stations are useful for air quality and soil moisture. The representability radius is found as a difficult concept by users. In terms of data accessibility, a majority of users still demand simple ways of downloading the data as CSV and NetCDF files.

The requirements' value goes beyond the pure goal of collecting them, as it makes data discovery more efficient and also allows clustering user's requirements, which enables identifying the complete or partial gaps, and finally could contribute to develop a global efficient strategy for in situ data measuring each EV driven by the user's needs.

Each advanced query in the GEOSS portal could be considered a parameterized user requirement. However, the GEOSS portal does not provide the necessary filters to allow for an automatic connection to the user requirements in G-reqs. In particular, the advanced search does not even provide a way to separate in situ data from the rest, and there is no way to query by spatial resolution, essential variable, thematic uncertainty, update frequency, or any of the three specific characteristics for in situ data: even distribution, coordinated measures, or representability radius. This prevents G-reqs from automatically checking for an existing dataset fulfilling the parameters of the requirement. In addition, if the GEOSS portal was able to store the advanced queries made by users, we will be able to generate a list of data requirements that could automatically populate G-reqs and could be used to rank the GEOSS portal results.

GEO is the right place to promote the G-reqs data model and methodology to gather user requirements, coordinate an effort among the thematic activities, generate consensus

on in situ products that are needed by the community, and find a way to make them accessible or create them. GEO could incorporate G-reqs as part of the GEOSS platform. This will allow users to better find the in situ products that fulfill their requirements but also to make visible that some demanded products are not available or do not yet exist and should be developed by the community in the future. With G-reqs, the GEO community should offer a new capacity or a service: to look for datasets that match some user data requirement but are not accessible now, and explore which community or network could be interested in producing a new in situ dataset to cover an emerging user's need.

References

1. Di, L. Distributed geospatial information services-architectures, standards, and research issues. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2004**, *35*, 187–193.
2. Heldens, W.; Burmeister, C.; Kanani-Sühring, F.; Maronga, B.; Pavlik, D.; Sühring, M.; Zeidler, J.; Esch, T. Geospatial input data for the PALM model system 6.0: Model requirements, data sources and processing. *Geosci. Model Dev.* **2020**, *13*, 5833–5873. [CrossRef]
3. Hu, L.; Yue, P.; Zhang, M.; Gong, J.; Jiang, L.; Zhang, X. Task-oriented Sensor Web data processing for environmental monitoring. *Earth Sci. Inform.* **2015**, *8*, 511–525. [CrossRef]
4. Wiegand, N.; García, C. A task-based ontology approach to automate geospatial data retrieval. *Trans. GIS* **2007**, *11*, 355–376. [CrossRef]
5. Li, M.; Guo, W.; Duan, L.; Zhu, X. A case-based reasoning approach for task-driven spatial–temporally aware geospatial data discovery through geoportals. *Int. J. Digit. Earth* **2017**, *10*, 1146–1165. [CrossRef]
6. Lehmann, A.; Masò, J.; Nativi, S.; Giuliani, G. Towards integrated essential variables for sustainability. *Int. J. Digit. Earth* **2020**, *13*, 158–165. [CrossRef]
7. Reyers, B.; Stafford-Smith, M.; Erb, K.H.; Scholes, R.J.; Selomane, O. Essential variables help to focus sustainable development goals monitoring. *Curr. Opin. Environ. Sustain.* **2017**, *26*, 97–105. [CrossRef]
8. Kissling, W.D.; Ahumada, J.A.; Bowser, A.; Fernandez, M.; Fernández, N.; García, E.A.; Guralnick, R.P.; Isaac, N.J.; Kelling, S.; Los, W.; et al. Building essential biodiversity variables (EBVs) of species distribution and abundance at a global scale. *Biol. Rev.* **2018**, *93*, 600–625. [CrossRef]
9. World Meteorological Organization; The 2022 GCOS ECVs Requirements. 2022. Available online: https://library.wmo.int/doc_num.php?explnum_id=11318 (accessed on 5 February 2023).
10. WMO. Requirements for Observational Data: The Rolling Review of Requirements. v11. 2018. Available online: https://space.oscar.wmo.int/observingrequirements (accessed on 5 February 2023).
11. Pulvirenti, L.; Parodi, A.; Lagasio, M.; Pierdicca, N.; Venuti, G.; Realini, E.; Gatti, A.; Barindelli, S.; Passera, E.; Rommen, B. Incorporating Sentinel-derived products into numerical weather models: The ESA STEAM project. In *Active and Passive Microwave Remote Sensing for Environmental Monitoring II*; SPIE: Philadelphia, PA, USA, 2018; Volume 10788, pp. 18–26.
12. O'Connor, L. NOAA's observing requirements collection process—Making a global difference. In *21st International Conference on Interactive Information Processing Systems*; AMS: Providence, RI, USA, 2005.

13. Cantrell, L.; Helms, D. The Continued Evolution of NOAA's Observing System Investment Assessment Process. In Proceedings of the American Meteorological Society 2016, New Orleans, LA, USA, 10–14 January 2016. Available online: https://ams.confex.com/ams/96Annual/webprogram/Paper281927.html (accessed on 5 February 2023).

14. Gallagher, F.; Griffin, V.; Saleh, R.; Li, X.; Marley, S.; Turner, S.; George, N.; Rangachar, R.; Covert, M.; Kurucz, P.; et al. Architecture Study for the Evolution of the NESDIS Ground Enterprise. In Proceedings of the AGU Fall Meeting 2021, New Orleans, LA, USA, 13–17 December 2021; Volume 2021, p. IN21B-01.

15. EEA. Copernicus In Situ Component Information System User Guide. 2022. Available online: https://cis2.eea.europa.eu/docs/guide.html (accessed on 15 February 2023).

16. Lush, V.; Bastin, L.; Lumsden, J. Geospatial data quality indicators. In Proceedings of the 10th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, Florianopolis, SC, Brazil, 10–13 July 2012.

17. GEO. Canberra Declaration. 2019. Available online: https://earthobservations.org/canberra_declaration.php (accessed on 1 February 2023).

18. GEO. Earth Observations in Support of the 2030 Sustainable Development. Document Version 1.1. 2017. Available online: https://www.earthobservations.org/documents/publications/201703_geo_eo_for_2030_agenda.pdf (accessed on 5 February 2023).

19. Anderson, K.; Ryan, B.; Sonntag, W.; Kavvada, A.; Friedl, L. Earth observation in service of the 2030 Agenda for Sustainable Development. *Geo-Spat. Inf. Sci.* **2017**, *20*, 77–96. [CrossRef]

20. OMG Standards Development Organization. Unified Modeling Language, Version 2.5.1. 2017. Available online: https://www.omg.org/spec/UML (accessed on 15 January 2023).

21. Tóth, G.; Jones, A.; Montanarella, L. The LUCAS topsoil database and derived information on the regional variability of cropland topsoil properties in the European Union. *Environ. Monit. Assess.* **2013**, *185*, 7409–7425. [CrossRef]

22. White, T.; Brantley, S.; Banwart, S.; Chorover, J.; Dietrich, W.; Derry, L.; Lohse, K.; Anderson, S.; Aufdendkampe, A.; Bales, R.; et al. The role of critical zone observatories in critical zone science. In *Developments in Earth Surface Processes*; Elsevier: Amsterdam, The Netherlands, 2015; Volume 19, pp. 15–78.

23. Brantley, S.L.; McDowell, W.H.; Dietrich, W.E.; White, T.S.; Kumar, P.; Anderson, S.P.; Chorover, J.; Lohse, K.A.; Bales, R.C.; Richter, D.D.; et al. Designing a network of critical zone observatories to explore the living skin of the terrestrial Earth. *Earth Surf. Dyn.* **2017**, *5*, 841–860. [CrossRef]

24. Domingo-Marimon, C.; Masó, J.; Prat, E.; Zabala, A.; Serral, I.; Batalla, M.; Ninyerola, M.; Cristóbal, J. Aligning citizen science and remote sensing phenology observations to characterize climate change impact on vegetation. *Environ. Res. Lett.* **2022**, *17*, 085007. [CrossRef]

25. GEO Data Working Group. GEO Data Sharing and Data Management Principles. GEO Knowledge Hub. 2022. Available online: https://gkhub.earthobservations.org/packages/pxdag-hq931 (accessed on 5 February 2023).

26. Portele, C.; Vretanos, P.A.; Heazel, C. OGC API Features—Part 1: Core. OGC 17-069r4. 2022. Available online: https://docs.opengeospatial.org/is/17-069r4/17-069r4.html (accessed on 5 February 2023).

27. GEO. Report on Usage of the GEOSS Platform. 2021. Available online: https://www.earthobservations.org/documents/pb/me_202105/PB-20-09_ReportonUsageoftheGEOSSPlatform.pdf (accessed on 5 February 2023).