2021

AN INTEROPERABILITY STRATEGY FOR THE NEXT GENERATION OF SEEA ACCOUNTING

Table of Contents

1. The current state of data interoperability in SEEA & a vision for the near future

SEEA requires the integration of substantial and diverse data. These include geospatial and other data not originally designed for statistical purposes, whose use is necessary for spatial modeling, which has proven challenging for some National Statistics Offices (NSOs) to implement. A variety of ecosystem service modeling platforms have been built over the last 15 years to meet various user demands,[1] as have numerous data viewers and dashboards, but their development has been uncoordinated.

These platforms often duplicate efforts, rely on data that are siloed, and rarely effectively reuse the knowledge gained from past modeling efforts. This document addresses the various challenges surrounding SEEA data and model interoperability, and a strategy for overcoming these interoperability challenges, with the audience being NSOs, data providers, and modelers. It draws on the experience of the Artificial Intelligence for Environment & Sustainability (ARIES) Project, which provides a feasible approach to make scientific data and models produced by diverse groups interoperable, rather than prescribing a particular modeling approach for each ecosystem service.[2]

A variety of initiatives have emerged in recent years that propose paths forward from the status quo. For example, the FAIR Principles propose steps to make data Findable, Accessible, Interoperable, and Reusable by both human users and by computers, which could automate modeling workflows. Interoperability is defined as "the ability of data or tools from non-cooperating resources to integrate or work together with minimal effort."[3] While the Open Science movement has succeeded in making data and code more Findable and Accessible through public data and code repositories, such repositories still typically struggle to achieve Interoperability and Reusability, and the global statistical community has recognized this challenge.

Following production of a 2018 guide on data interoperability for the development sector,[4] the UNSD Working Group on Open Data has added a workstream on data interoperability,[5] the U.N. Global Statistical Geospatial Framework has identified statistical and geospatial interoperability as one of its five Principles,[6] and the Global Group on Earth Observations' Earth Observation for Ecosystem Accounting (EO4EA) Initiative has begun development of a consensus document on "Ecosystem Accounts-Ready Data" that acknowledges a key role for the FAIR Principles.[7] Finally, data cubes, defined by OECD as a "multi-dimensional structure for storing statistical information" have gained popularity in the statistical community as a more consistently structured way to store information, including spatial data.[8] Like most open-data efforts, however, data cubes have generally struggled to achieve interoperability and

[1] https://seea.un.org/ecosystem-accounting/biophysical-modelling
[2] https://aries.integratedmodelling.org/. For a detailed technical description of the software and semantic modeling approach used by ARIES, see https://docs.integratedmodelling.org/technote/.
[3] Wilkinson et al. 2016; https://www.nature.com/articles/sdata201618.
[4] https://www.data4sdgs.org/resources/interoperability-practitioners-guide-joining-data-development-sector
[5] https://unstats.un.org/open-data/
[6] http://ggim.un.org/meetings/GGIM-committee/9th-Session/documents/The_GSGF.pdf
[7] https://www.eo4ea.org/
[8] Data cubes describe spatiotemporal data well, particularly in times series. They consist of a multi-dimensional array of values, associated with the data. The data represent some measure of interest, and each dimension corresponds to a separate measurement associated with the same data.

reusability. While these initiatives suggest a high level of interest in improving data interoperability, the community has not yet coalesced around a solution to the critical problem of interoperability.

In 2020, UNSD began work on a concrete path toward data and model interoperability to support the compilation of SEEA Ecosystem Accounting (SEEA EA), collaborating with the ARIES team to develop an approach, ARIES for SEEA, that can produce SEEA accounts anywhere on Earth. ARIES combines the use of global data and generalized modeling approaches in data-scarce countries yet is simultaneously able to customize accounts wherever improved data and model resources exist – thus meeting the needs of both countries with limited data who want to compile initial accounts and those capable of more highly customized SEEA EA approaches.[9]

Through over a decade in development, ARIES has long supported FAIR modeling approaches by building a *semantic web*[10] of data and spatial models that achieve high-level *semantic interoperability* (which enables a receiving system to properly understand the meaning of data that are exchanged, reusing it in an appropriate manner, as opposed to lower-level *syntactic interoperability*, which relies on the use of compatible data formats and communication protocols).

ARIES makes a large and growing collection of data and models easily accessible to users with limited experience in spatial modeling, including NSOs, while simultaneously ensuring *appropriate reuse* of models, as conditions for model reuse are explicitly encoded, guiding the selection of the most appropriate approaches for the time, place, and scale of the analysis. ARIES can also facilitate reporting on key global initiatives such as the Sustainable Development Goals, Post-2020 Global Biodiversity Framework, Paris Climate Agreement. Critically, the approach offers NSOs a starting point to begin needed conversations with data producers and modelers, by building basic initial estimates based on global data and preexisting models then continually working with data producers and modelers to improve initial results using more local-scale scientific data and knowledge.

A long-term shared vision of UNSD, ARIES for SEEA, and EO4EA is that (1) all key data and models needed to compile SEEA accounts and related global indicators (e.g., SDGs, post-2020 Biodiversity Goals) are interoperable, while (2) researchers independently use the FAIR Principles when developing new data and models, making them seamlessly ingestible by interoperability-centered modeling approaches like ARIES (which simultaneously provide a high degree of transparency through automated production of reports and provenance information – a "digital signature" of which data and models were used).

Countries with limited data and technical capacity can benefit strongly from an interoperability-focused approach, by gaining access to context-appropriate data and models that can be properly assembled by computers. For NSOs and researchers in technically advanced nations, an interoperability focus offers a way to diffuse scientific knowledge (and recognition for their modeling efforts) more rapidly through greater (appropriate) reuse of their data and models.

The practical outcome of this vision is that SEEA accounts, ecosystem service assessments, and related indicators will be (1) rapidly recompilable as new science emerges, (2) quickly produced to show the most recent trends as new annual data become available, with (3) robust international comparisons made possible by common global data, while country-specific customization is still easily done. This vision moves high-quality, meaningful statistical information from scientists into the hands of decision

[9] http://aries.integratedmodelling.org/aries-for-seea-eea-for-rapid-natural-capital-accounts-generation/

[10] A "web of data" interlinked so that both people and computers could traverse across databases over the network

makers, the public, and the media as quickly as possible. Further discussions on this strategy are needed to develop common understanding around interoperability and its benefits and consensus around any areas of disagreement, turning conceptual buy-in into widespread implementation.

2. Roles and responsibilities in achieving interoperability

The strategy described below develops a concrete approach to move toward ambitious yet achievable goals for interoperability in the SEEA community, and describes the roles and responsibilities of various stakeholders working in the geospatial modeling community (e.g., data providers, modelers, platform hosts) to achieve extensive use of SEEA by countries around the world. By working together as a network of networks, these various groups can support successful and widespread implementation of the SEEA.

In addition to the individually held roles and responsibilities for the groups noted below, quality control (including the inclusion of recommendations about the appropriateness of data/model reuse for different use cases) is a shared goal of all who contribute and use data. Quality control is an "all hands on deck" exercise, and the topic will be discussed further in a companion ARIES Technical Note. The ARIES Project implements these strategies, and seeks partner organizations with an interest in following these principles to maximize data and model interoperability and reusability in a decentralized, networked fashion.

2.1 Proposed roles & responsibilities of data providers

Interoperability for both people and computers is achieved not just by making data public on an independent website or even in established repositories (i.e., today's typical open-science practices), but by using common and established data formats, hosting protocols, and semantics.[11] This guarantees that data achieve higher-level semantic interoperability, rather than just syntactic interoperability. To maximize the interoperability of data assets, data providers can:

a. Whenever possible, expose and maintain key spatial datasets as Open Geospatial Consortium services[12] using networked infrastructure (e.g. GeoServer, PostGIS[13]) - hosted independently, through the U.N. Global Platform, or through or other networks explicitly designed for semantic interoperability.

b. Plan for data compatibility using open, widely available standards[14] and ensuring that associated metadata are complete, correct, and semantically meaningful. Host tabular data in machine-accessible formats and provide an Application Programming Interface (API) for access whenever possible.

[11] Semantics define concepts and the relationships between them, formalizing the meaning of underlying data and models in a meaningful way for both people and computers.
[12] https://www.ogc.org/standards
[13] http://geoserver.org/about/ https://postgis.net/
[14] E.g., NetCDF, Hierarchical Data Format (.hdf), Cloud Optimized Geotiffs

c. Produce Uniform Resource Name (URN)-specified, non-semantic resources from each dataset of interest and publish to a networked node to enable its semantic annotation by any participant.

To be semantically interoperable, hosted data must carry consistent, clear semantic meaning that can be understood by people and computers on a network, by:

a. Committing to use a common set of ontologies and vocabularies[15] and enlisting partner institutions as users and contributors to ensure that semantics are developed collaboratively. The use of common semantics for all data and model elements ensures that both people and machines are certain which data and model components are interchangeable, and which are not.

b. Identifying a single point of contact for each data-providing institution to follow the semantics development and tooling efforts and be responsible for the consistent use of the vocabulary (to describe all relevant data and model components) and growth of the vocabulary (when data or models describing new concepts are introduced).

c. Together, a larger semantics community must gradually move the task of semantic annotation to data producers. This will require the production of best practice documents, handbooks for specific problem areas, and ad-hoc tooling.

2.2 Proposed roles and responsibilities of modelers

For modelers, semantic interoperability requires several adjustments to typical model design and philosophy, which require up-front work. This investment pays off quickly by enabling much more frequent reuse of past code (e.g., when models can be more easily linked together, with one model automatically calling on another to produce a needed input), building more quickly on others' properly credited work while requiring less development of new code by reusing existing code, and taking full advantage of the modularity and flexibility offered by this approach (Box 1).[16] Modelers interested in coding with an eye toward interoperability can:

a. Adopt design principles and guidelines for independently produced, interoperable model *projects* using distributed version control software.[17] Scope projects carefully to bring models through the stages of development from experimental/locally hosted, through project-based/institution hosted, to public/globally available.

b. Adopt a more modular, less monolithic model design process. Each model or data annotation should describe a single concept. This facilitates interconnection to other models, and allows dependencies to be managed through scoping rules instead of hard-coded linkages.

c. Learn the importance of tracking provenance[18] of all official products obtained through modelling. Annotate data and models to maximize the value of provenance information, based on best practices.

[15] E.g., https://f1000research.com/articles/6-686, https://sdmx.org/, https://www.xbrl.org/
[16] e.g., https://www.biorxiv.org/content/10.1101/2021.02.23.432363v1.abstract
[17] https://git-scm.com/book/en/v2/Getting-Started-About-Version-Control
[18] A description of the data or model source, with connections to its full metadata.

d. As a community, develop strategies and an incentive structure to overcome the status quo of non-interoperable model development.

Additional approaches for making existing scientific models interoperable within the ARIES Network are described in this technical note.

Box 1. The benefits of interoperability: An example for sediment retention accounting

Sediment retention is an ecosystem service frequently included in ecosystem accounting and ecosystem service assessments, for which biophysical models are typically required. These models rely on approaches like the Revised Universal Soil Loss Equation (RUSLE), implemented by well-known platforms such as InVEST, ARIES, and LUCI as well as numerous one-off studies. RUSLE requires several spatial data inputs such as a digital elevation model (DEM), land use-land cover, soils, and rainfall erosivity data, plus coefficients related to land-use and agricultural practices.

A typical modeling exercise begins by searching for the best available spatial and tabular data about agricultural practices (ideally using national data but relying on global data where local sources are unavailable). Spatial data collection requires GIS expertise to find, download, and prepare the needed data, a process that can take days to a few weeks to accomplish depending on the analyst's skill and experience. To find appropriate tabular data, a researcher familiar with the soil erosion literature must search for and read relevant literature for their study area, then apply expert judgment to determine the best-available model parameters for their context. This process also typically requires weeks of work (at minimum), and the researcher never knows with certainty when the job is done – there is always the possibility that other, undiscovered, data sources exist. After completing their study, the researchers may (or may not) publish their data in a public repository, and transparently report their model coefficients and rationale for their selection in publicly accessible literature for future reuse.

Using an interoperability-focused approach, these same researchers would, at the end of their project, (1) place their data in a public repository using machine-actionable formats and with semantically meaningful metadata and (2) publish any new code – in this case including lookup tables – in a public repository, with semantically meaningful metadata and a description of data quality and appropriate reuse conditions for this knowledge (e.g., within a given ecosystem types or other spatial extent, or spatiotemporal scale), as described in this interoperability strategy.

By doing so, these researchers would make their data both easily *interoperable and reusable.* A future assessment conducted in the same country (or a neighboring country with similar enough conditions) could automatically reuse knowledge from the earlier study – integrating data for new assessments in a matter of hours rather than weeks to months. Time spent on GIS processing and hunting for parameters could instead be spent on other technical work and stakeholder outreach. As more researchers contribute data in an interoperable fashion, scientists would reuse and improve on past data more effectively than is possible through time-consuming and often incomplete literature searches. Further, as more and more *models* are made interoperable (including alternative approaches to RUSLE), the most suitable one can be chosen for the problem at hand. Today's best-case scenario for a modeler is a decision tree or database (e.g., https://esdac.jrc.ec.europa.eu/content/global-applications-soil-erosion-modelling-tracker) to guide the modeler toward the best approach, after which they must develop data and parameterize and run the chosen model. An interoperability approach allows such decision making about data and model selection to be automated and delegated to faster and more robust artificial intelligence (AI) algorithms, with a high degree of transparency, e.g., through automated reporting and provenance systems built into ARIES. This automated reporting process provides a "digital signature" of each analysis across the full knowledge chain.

A widely adopted interoperability strategy provides several tangible benefits – it makes modeling faster, more efficiently reuses past scientific knowledge, and ensures that data and models are reused appropriately. With wider use of interoperability-focused practices and quality control of data entering the shared knowledge base, quality and speed of future ecosystem accounts can both be ensured (see the vision statement in bold at the end of Section 1 of this strategy).

2.3 Proposed roles and responsibilities of hosting institutions

The medium- to long-term commitment of hosting institutions (e.g., existing data repositories, large research groups, or NSOs) is fundamental for interoperability to persist over time. Proper incentives are needed so that hosts maintain resources in an interoperable manner. Institutions may wish to formally host data and models. A possibility would be to make results available through an API or to contribute them to the ARIES Network by setting up a k.LAB node (which offers ways to access data and models through both ARIES and other approaches)[19].

Doing so will involve software, hardware, and personnel needs. In case of setting up a node, support will be available from the ARIES development team in the form of training material, responses to frequently asked questions, and troubleshooting documentation. A fully interoperable data and model system moves away from the paradigm of a centralized entity that provides and hosts all the needed resources, to a community of hosting members, within a peer-to-peer system. An ecosystem of multiple nodes provides a more stable, flexible and powerful network, which concurrently offers each member full ownership and control of critical data and models that are not meant to be shared with the wider community.

3. Putting the strategy into practice

To put this interoperability strategy into practice, we envisage four steps:

1. **Pilot testing.** Pilot testing will be a critical part of broadening the interoperability strategy to a wider community – building a larger group who understands interoperability on a conceptual level, the benefits it offers, and that will advocate for its wider use. Pilot testing of the interoperability strategy should be conducted with selected partners in order to work through unresolved issues, smooth out communication, technical, or strategic issues, and keep an eye toward scaling up to create a global network of data and models. From the strategic perspective, it is important to set priorities that can produce tangible benefits with the addition of each new partner or knowledge area (i.e., engaging a very small number of strategic partners first, and building from there). A plan for scaled expansion should proceed in a manner where each new added partner can demonstrate a multiplicative effect, rather than an additive one, where newly added data and models support an increasingly large number of useful outputs.

2. **Engaging key stakeholders.** From the data and model providers' perspective, the most important task is identifying the *key areas of knowledge*, and the corresponding *communities of practice*, where a semantic annotation and interoperability effort should be undertaken. Natural candidates could be the U.N.-led communities in charge of the wider SEEA Central Framework accounts not currently covered by ARIES for SEEA (e.g., energy, air emissions, agriculture-forestry-fisheries), selected communities within the European Environment Agency, Joint Research Center, the European Space Agency, NASA, the SEEA EA Technical Committee, Natural Capital Project, the EO4EA GEO Initiative that is invested in SEEA, and other GEO initiatives that address specific problem areas such as GEOBON for biodiversity.

[19] The k.LAB Node is the server infrastructure that provides knowledge for the k.LAB network (which runs ARIES). Its main purpose is the distribution of knowledge for modelling engines: providing data and computed models from URNs, hosting worldviews, software components and semantic projects.

3. **Governance.** A number of institutions and NSOs worldwide (including the European Environment Agency, European Space Agency and Interamerican Development Bank) have shown interest in further testing or integration of their data and models following demonstrations of the ARIES for SEEA application. There is thus clear interest in working toward greater interoperability. It is proposed that (1) the UN Committee of Expert on Environmental-Economic Accounting (UNCEEA) could play a coordinating role, working together with identified stakeholders and other groups such as the Green Growth Knowledge Platform (GGKP) Natural Capital Data working group, and/or (2) a new group could be established, with a UNSD-provided Secretariat and co-chair(s) from the NSO and scientific/geospatial communities. NSOs, science/geospatial experts and agency representatives, and academics can serve as ambassadors for this effort in their respective communities.

4. **Training and capacity building.** There is a need for further training and instructional materials for data providers, modelers, and platform hosts that specify steps to be taken to move towards interoperability. This should include continued outreach on the value of interoperability, how to most effectively reuse interoperable data and models, and how to contribute data and models to an interoperable ecosystem.

4. Conclusions

The success of an interoperability strategy depends on two things: mature technology and the willingness of a large community to adopt it. With the development of ARIES for SEEA, a proof-of-concept approach for widespread semantic interoperability of data and models is available for much wider use. This can benefit both compilers of SEEA accounts and related indicators and scientists who produce data and models, who could contribute their knowledge to a broad and interoperable framework, improving knowledge production that better informs decisions.

Adoption and implementation of a shared interoperability strategy is extremely important to advancing and scaling up the implementation of SEEA EA in countries. Coordination on an interoperability strategy has initial costs and a learning curve needed to build a common vision while encouraging individual scientists to better organize, clean, and integrate the information needed to compile SEEA accounts. However, by improving knowledge reuse (i.e., reducing reinvention of the wheel) and speeding accounts compilation, a robust interoperability strategy has a very strong return on investment – substantially improving national and global capacity to produce timely and decision-relevant information for SEEA and other critical global indicators. With the technology available to solve the well-acknowledged interoperability problem and the stakes high for timely and informed environmental decision making, it is urgent that the global data, modeling, and statistical communities involved in development and implementation of the SEEA EA develop, endorse, and begin to move forward on an inclusive and shared interoperability strategy.

The strategy was developed as part of the EU funded project Natural Capital Accounting and Valuation of Ecosystem Services (NCAVES)

CONTACT

ARIES: ARtificial Intelligence for Environment & Sustainability

aries.integratedmodelling.org

aries@integratedmodelling.org