remote sensing

MDPI

Article

The Austrian Semantic EO Data Cube Infrastructure

Martin Sudmanns [1,*], Hannah Augustin [1], Lucas van der Meer [1], Andrea Baraldi [2] and Dirk Tiede [1]

[1] Department of Geoinformatics—Z_GIS, University of Salzburg, 5020 Salzburg, Austria; hannah.augustin@plus.ac.at (H.A.); lucas.vandermeer@plus.ac.at (L.v.d.M.); dirk.tiede@plus.ac.at (D.T.)
[2] Spatial Services GmbH, 5020 Salzburg, Austria; andrea6311@gmail.com
* Correspondence: martin.sudmanns@plus.ac.at; Tel.: +43-(0)-662-8044-7584

Abstract: Big optical Earth observation (EO) data analytics usually start from numerical, sub-symbolic reflectance values that lack inherent semantic information (meaning) and require interpretation. However, interpretation is an ill-posed problem that is difficult for many users to solve. Our semantic EO data cube architecture aims to implement computer vision in EO data cubes as an explainable artificial intelligence approach. Automatic semantic enrichment provides semi-symbolic spectral categories for all observations as an initial interpretation of color information. Users graphically create knowledge-based semantic models in a convergence-of-evidence approach, where color information is modelled a-priori as one property of semantic concepts, such as land cover entities. This differs from other approaches that do not use a-priori knowledge and assume a direct 1:1 relationship between reflectance values and land cover. The semantic models are explainable, transferable, reusable, and users can share them in a knowledgebase. We provide insights into our web-based architecture, called Sen2Cube.at, including semantic enrichment, data models, knowledge engineering, semantic querying, and the graphical user interface. Our implemented prototype uses all Sentinel-2 MSI images covering Austria; however, the approach is transferable to other geographical regions and sensors. We demonstrate that explainable, knowledge-based big EO data analysis is possible via graphical semantic querying in EO data cubes.

Keywords: semantic earth observation data cube; big earth observation data; sematic analysis; explainable artificial intelligence; semantic enrichment; scalable system architecture

Citation: Sudmanns, M.; Augustin, H.; van der Meer, L.; Baraldi, A.; Tiede, D. The Austrian Semantic EO Data Cube Infrastructure. *Remote Sens.* **2021**, *13*, 4807. https://doi.org/10.3390/rs13234807

Academic Editor: Pieter Kempeneers

Received: 30 September 2021
Accepted: 24 November 2021
Published: 26 November 2021

1. Introduction

Our concept and implementation of a semantic Earth observation (EO) data cube architecture aims to facilitate semantic queries by including computer vision (CV) directly at the EO data cube level. The EO data cube is a concept and tool for managing spatio-temporal big EO data that aims to lower barriers to EO data access [1]. Data provision has also evolved with the establishment of Analysis-Ready-Data (ARD), such as the CEOS ARD for Land (CARD4L [2]), which facilitates transferrable yet hard-coded and application-specific algorithms (e.g., water observations from space (WOfS) [3]). Despite these technical and organisational improvements, users still have to work with reflectance values, which lack semantics or inherent meaning. Reflectance values are sub-symbolic numerical variables [4] that represent the proportion of measured radiation striking the (Earth's) surface to the radiation reflected from the surface material. These reflectance values (color of the material) are not usually uniquely associated to land cover entities; therefore, they require users to create data-driven workflows to produce information from these values for each application anew. This poses a big challenge for many end-users of big EO data, especially now that the amount and diversity of EO data users is increasing, including more non-expert users [5,6]. Our architecture combines an artificial-intelligence-based (AI-based) CV system with EO data cubes for optical satellite images so that end-users can perform sophisticated analyses without requiring programming or advanced technical skills.

In addition to cloud and HPC technology, the number of interactive web-based platforms with programming interfaces to EO data has recently increased and open the potential to a paradigm shift from downloading and processing data locally to web or cloud-based processing [7]. Examples include Jupyter notebooks, the Google Earth Engine code editor, and the Web coverage processing service (WCPS). Jupyter notebooks are a development environment that provides text descriptions together with code to allow comprehensive and reproducible workflows [8]. The WCPS is an Open Geospatial Consortium (OGC) standard for declarative queries of coverages (e.g., raster data, images) [9]. Google Earth Engine (GEE) is a web-based EO data analysis platform that supports multiple programming languages [10].

Regardless of these technical developments, a significant proportion of big Earth data users, including but not limited to EO data, still do not use tools with web-based code editors [5]. These tools generally require specialized skills (e.g., procedural or declarative programming) and demand significant time investment to learn how to access and handle data on each platform [5]. Implementations with programming-free graphical user interfaces (GUIs) exist, but typically are either limited in their analytical capabilities in exchange for requiring fewer technical skills (e.g., simple map viewers), or are tailored to specific and pre-defined applications and outputs, such as software systems for agricultural monitoring (e.g., sen2agri [11,12], Sen4CAP [12]). Figure 1 shows how some web-based EO data analysis approaches address this challenge according to the dimensions of usability, features, variety of target users, and variety of applications. It is an ongoing challenge to develop web-based approaches for EO image analyses that are feature-rich while still being application independent and easy to use by a variety of users [13].

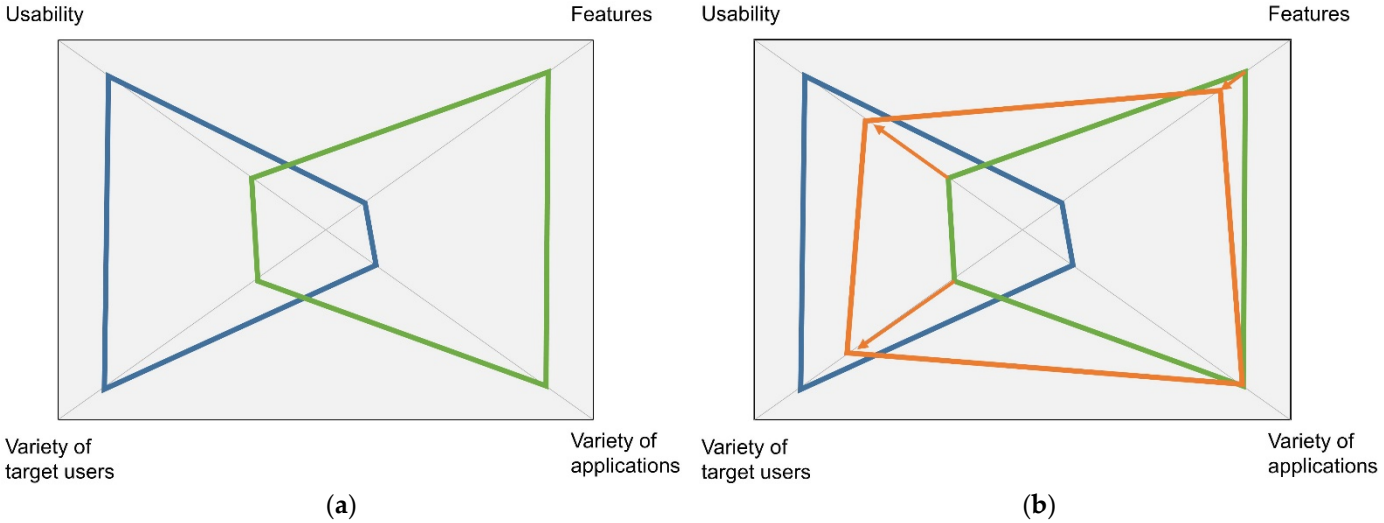

Figure 1. Simplified spider diagrams compare different approaches for web-based EO data analyses (**a**): systems and interfaces that are easy to use and target various users are less transferable to general applications and provide fewer features (blue polygon), e.g., map viewers. Systems and interfaces with high applicability and many features are more difficult to use and require programming skills, limiting the target users (green polygon), e.g., web-based code editors. (**b**) It is an ongoing research gap to provide better usability for a larger target group while sacrificing as few features as possible and being generally applicable to various application domains (orange polygon). Specialized cloud services, such as the Copernicus Data and Information Access Services (DIAS), which provide virtual machines (VM) that need to be managed by users themselves, are considered here as a different layer in the technology stack.

The semantic analysis of EO images is an approach to shift workflows from technical EO considerations to users' domain knowledge. It relies on image understanding (vision) capabilities to bridge the gap between the observations and semantic concepts. Reconstructing the four-dimensional physical world (i.e., scene) from two-dimensional images (i.e., scene-from-image reconstruction), is an inherently ill-posed problem for two

main reasons. First, a two-dimensional image is reduced in its dimensions compared to the four-dimensional physical world. The dimensionality reduction is inherent to the image acquisition process, although the temporal dimension can be somewhat recovered by frequent acquisitions. Second, an optical EO image is composed of (regularly) sampled reflectance values (i.e., pixels) from the land cover material (e.g., soil, leaves, rock). Therefore, the pixel-based spectral signatures of the reflectance values may be associated to several land cover types, variances of their state or condition, or represent a mixture of land cover types. Reflectance values are necessary but not sufficient to reconstruct entities from images because these values alone generally cannot be directly mapped to land cover in a 1:1 relationship. In some cases, reflectance values may not even be related to land cover, such as clouds or smoke. Examples for both cases are dark shadow and water, which can have similar spectral signatures [14]. Additional information is necessary to solve the ill-posed problem of identifying land cover from EO images. The spectral signature (i.e., color) is only one attribute and additional information needs to be added, e.g., in a convergence-of-evidence approach. Further information could include the temporal trajectory, pixel context or spatial object information (e.g., size, texture, shape).

CV aims to produce automated scene-from-image reconstruction using optical (EO) images and presents an opportunity to implement semantic approaches in big EO data analyses. These analyses allow users to formulate queries using semantic concepts instead of writing low-level procedural or declarative code, which starts with reflectance values. CV approaches that converge multiple sources of evidence can produce information on a big EO data scale (e.g., national information layers) in a systematic, automated, explainable and computer-based way.

Image understating may be achieved by human interpretation when using a small number of images. However, using EO data cubes to scale big EO data analyses to a national level allows more users to make evidence-based decisions. Information produced from EO data cubes can contribute to monitoring sustainable development goals (SDGs) and improving the understanding of (natural) systems with essential variables (EVs). In automated big EO data analyses involving thousands of images (or even more), the meta quality indicators of products and processes becomes increasingly important. Some of these indicators are not easy to achieve and obtain, e.g., the timeliness and frequency of products, or the transferability and reproducibility of methods. These requirements increasingly exclude the option to perform these types of analyses on a single person's computer or with manual preparation and interaction. Questions around the distribution, interoperability, and interactivity of the results raise concerns about whether a rendered map on a (printed) journal publication or report is still sufficient. The EO data cube can be a flexible central data and information space to account for the shifts of workflows and requirements.

Sen2Cube.at is a first step towards a CV-based EO data cube approach and named after its first prototypical implementation, "Sentinel-2 Semantic Data Cube Austria" (https://sen2cube.at, accessed on 28 October 2021). It implements a semantic EO data cube [15] in a state-of-the-art cloud-based environment. Our approach automates the semantic enrichment on a big EO data scale and demonstrates that semantic EO data cubes can be scaled to at least national levels. Moreover, the interactive web-based GUI allows users, whether EO-novices or experts, to conduct analyses based on big EO imagery content using semantic querying without requiring programming. To the best of our knowledge, a similar approach has not yet been implemented or published before.

This paper connects our Sen2Cube.at architecture to related work, describes the system details, including an initial prototypical use-case for Austria, and shares the associated challenges and the direction of future research. Section 2 contextualizes our approach to big EO data management and semantic analyses. Section 3 lays out the system's requirements, highlighting the semantic enrichment of optical images and EO data cubes as the data management backbone. Section 4 explains the technical components of the architecture. Section 5 illustrates the architecture's application in the Austrian national semantic EO data cube infrastructure. Section 6 discusses the advantages and limitations, and Section 7

shares our conclusions and outlook as well as a variety of other potential use-cases beyond using Sentinel-2 data covering Austria.

2. Related Work

2.1. Big Earth Observation Data Management and Processing

Approaches to big EO data management and processing include operational EO data cubes on national and regional scales. Some of them provide a Python-based API that allows users to load data as Python xarray objects [16]. These include the Open Data Cube [17] (see implementations in Digital Earth (DE) Australia [18], DE Africa, Switzerland [19,20], Colombia [21], Vietnam [22], Armenia [23], and Catalonia [24]) and the xCube, implemented in the Euro Data Cube [25] as part of the Euro Data Cube Facility. The Earth System Data Lab (ESDL) is a project to create multi-variate data cubes with a Python and Julia API [26]. Rasdaman [27] is an array database system with deployments in the EarthServer projects [28] and CODE:DE, in particular the BigDataCube service [29]. Other systems for big EO data analyses include Google Earth Engine [10], the JEODPP, which is a Joint-Research-Center-hosted (JRC) platform for EO analyses making use of Jupyter notebooks for batch processing as well as visualisations [30], the Big Earth Data Science Engineering Project (CASEarth) [31], and the SITS R package, which is behind the Brazil Data Cube [32]. An overview of data cubes is provided in [33], and a comprehensive overview of processing platforms is provided in [34].

2.2. Semantic Approaches in Big Earth Observation Analytics

Early approaches in remote sensing image data mining (e.g., [35]), semantic content based image retrieval (e.g., [36,37]), and ontology based approaches (e.g., [38,39]) argue for the use of semantic content in image querying and analysis, but are limited to image searches (i.e., no analysis) or small, application-specific areas-of-interest (AOI). The EO data domain has moved from data scarcity to a continuous data stream of observations allowing change detection based on time series analysis and a transition to monitoring approaches [40]. Many approaches and algorithms make use of dense EO time series [41–43], but most focus on the raw observations or derived indices as analysis input (see [32] arguing for time-first approaches) or are application-dependent. The authors of [13] highlight the need for a systematic, operational implementation of a knowledge-based system for EO analyses, and also demonstrate application-specific examples. To the best of our knowledge, no one has systematically conducted and implemented generic semantic enrichment (i.e., initial spectral categorization) of big EO data as application-independent categories into an EO data cube infrastructure for time-series and/or big EO data analysis as prototypically outlined by the authors in [44]. State-of-the-art tools for automated semantic enrichment usually generate lower semantic levels than a complete land cover classification such as the FAO LCCS (land cover classification system). Further enrichment requires specified rules in the spatial (e.g., object-based analyses) or temporal domain (e.g., time series analyses) in which the semantics of entity states and potential changes are defined as spatial, temporal and thematic attributes. These rules allow semantically enriching observations up to land cover types (e.g., FAO LCCS), as well as the identification of spatio-temporal trajectories of changes [45].

3. Conceptual Foundations and System Requirements

3.1. Earth Observation Data Cubes

An EO data cube is a way of organising EO data that abstracts data storage so that users can access EO data based on spatio-temporal coordinates rather than their file names or directory structures at a data centre [46,47]. EO data collected as individual image files located in directories with descriptive names on the hard drive is acquisition-oriented and straightforward but requires users to know filename conventions and internal directory structures of data centres. This makes time series analyses particularly cumbersome. However, pixels within an EO image have a position on Earth and an observation time,

which can be described by a coordinate tuple. Therefore, according to [48], an EO data cube is considered a multi-dimensional structure with at least one non-spatial dimension (e.g., time), where coordinate tuples of the dimensions are used for data access. Although a data cube can feature any number of dimensions, it is usually called a 'cube' [48]. The data might be still stored as individual image files with an index database and transformed into an in-memory multi-dimensional data structure on the fly or physically organized on the disk in a multi-dimensional structure to allow access that is more efficient. In all cases, the abstraction provided by an EO data cube allows users coordinate-based access as well as several multidimensional operations, regardless of how the data are named or stored. Still, some authors distinguish between image collections and data cubes [32].

If semantically enriched, EO data cubes are currently the most suitable technical foundation for our approach, despite being purely grid-based. A semantic EO data cube or a semantics-enabled EO data cube is defined as 'a data cube, where for each observation at least one nominal (i.e., categorical) interpretation is available and can be queried in the same instance' [15]. Here, the relationship between an observation and interpretation features a cardinality of 1:N, considering the reflectance values of all the bands as one observation. Vector-based storage would be additionally required to completely semantically-enable EO data [7,26].

3.2. Semantic Enrichment in the Earth Observation Domain

The big EO data domain requires automated approaches for generating information worldwide, but automated output currently produces high levels of semantic granularity (i.e., fewer categories) and is mostly application-dependent. The limitations are mainly due to the ill-posed problem that pixel-based spectral signatures can be associated with more than one land cover type. Therefore, the semantic enrichment of EO images refers to achieving a certain level of interpretation (i.e., mapping data to symbols that represent stable concepts). In the case of EO imagery represented in a grid, this means mapping each raster cell to an interpretation and represented as nominal or partially ordinal (e.g., vegetation categorized by increasing greenness or intensity) variable. The initial interpretation serves as input for classification and is a necessary preliminary step to CV.

Our architecture uses the SIAM (Satellite Image Automated Mapper) software as the initial, data-driven and bottom-up semantic enrichment [49–52]. SIAM automatically categorizes optical multi-spectral EO imagery from multiple sensors (e.g., Sentinel-2, Landsat-8, AVHRR, or very high spatial resolution sensors), given calibration to at least top-of-atmosphere (TOA) reflectance [50,53]. SIAM is considered automatic because it runs without user-defined parameterisation or training samples; instead, it relies on a-priori knowledge encoded in a decision tree that is based on a physical-spectral-model and applied to each pixel. The decision tree maps each calibrated observation to one stable, sensor-agnostic multi-spectral color name (i.e., category) based on its location in a multi-spectral reflectance hypercube. The result is a discrete and finite vocabulary of color names for observations that refer to hyper-polyhedra (e.g., multi-dimensional shape) within a multi-spectral feature space. The application-independent color names are mutually exclusive (i.e., every observation belongs to one and only one partition) and exhaustive (i.e., the entire multi-spectral reflectance hypercube is partitioned). These color names are semi-symbolic and should be understood as 'green-like vegetation' or 'blue-like water'. SIAM's relatively low-level, generic semantic enrichment has been independently validated up to a continental scale [50] and is capable of producing different granularities, from coarse (i.e., 18 color names) to fine (i.e., 96 color names), as well as additional data-derived information layers (e.g., multi-spectral greenness index, brightness).

3.3. Artificial-Intelligence-Based Expert System

An expert system (i.e., information processing system) typically infers information by applying expert rules from a knowledgebase to facts stored in a factbase [54]. Humans interact as users and as experts (potentially accompanied by knowledge engineers) adding their

knowledge to the system [54]. Expert knowledge can be summarized as a problem-solving strategy based on stored procedures, which can be applied to a task. Thus, knowledge engineering aims to translate human knowledge of the world or a domain subset into machine-readable rulesets, supported by an appropriate scheme such as ontologies.

Our proposed concept is an AI-based expert system for knowledge-based semantic EO analyses in a browser-based web application with user-definable (custom) graphical semantic models. The graphical semantic models represent general and specific expert knowledge of the world, formalized using a semantic querying language, and serve as the CV system's a-priori knowledge. The semantic models are stored in a knowledgebase and can be re-used or transferred to multiple AOIs and time periods. Every new semantic model incrementally augments the knowledgebase. The usability of Sen2Cube.at improves by sharing semantic models in the knowledgebase, which can be exchanged between expert and non-expert users. The factbase contains image data (i.e., reflectance values), semantically enriched information layers and any additional information layers (e.g., DEM, thematic information) and provides them as facts to the inference process. We use a semantically enriched EO data cube as a factbase because it facilitates spatio-temporal access to data in the cloud. The inference engine is the component of the system that processes the semantic models from the knowledgebase by translating them into queries against the factbase. Thus, in our approach, a semantic query refers to the execution of a semantic model using a spatio-temporal subset of the factbase, and an inference refers to the information production by means of semantic querying.

The information production in our system are conceptually broken into the following, consecutive steps:

1. Initial generic semantic enrichment: exclusive and exhaustive partitioning of spectral signatures into generally applicable color information. Users can apply them as the basic building blocks to define general semantic entities;
2. Convergence-of-evidence to increase the semantic granularity: semantic entities are defined using multiple sources of evidence, including color information, the temporal dimension, and potential auxiliary data and information, e.g., a digital elevation model (DEM);
3. Extraction of land use/land cover units using the semantic entities in specific applications for better-posed big EO data queries and analyses (e.g., semantic content-based image retrieval (SCBIR, a method to retrieve images based on a semantic description of their content), cloud-free compositing, automatic change detection).

3.4. General System Requirements

We defined the following general system requirements for a semantic EO data cube architecture:

(a) Application support
- Facilitate at least the four following use-cases: (a) SCBIR, (b) best-pixel selection for user-defined composites (e.g., cloud-free), (c) location-based querying (a.k.a. pixel/point drill), and (d) parcel-based querying (a.k.a polygon drill);

(b) Semantic EO data cubes and computing infrastructure
- Abstract data storage and access via EO data cubes, which contain at least one interpretation for every observation in space and time that can be queried in the same instance;
- Conduct automated EO data pre-processing, semantic enrichment and indexing; Utilize state-of-the-art cloud-based infrastructure (i.e., lightweight container-based virtualisation technology such as Docker) to reduce costs for installation and maintenance.

(c) User interaction and interfaces
- Process in the cloud, implementing the big data paradigm to 'bring user to the data, not data to the user';

- Require no user-side installation beyond a standard web browser and Internet connection;
- Abstract data access and algorithms via a graphical querying language and semantic models as a-priori information in an explainable AI approach;
- Allow interactive system use and batch processing;
- Provide a programming-language independent API;
- Facilitate multiple, concurrent users, including separate spaces for each user. Provide further information about how to use the system as a manual and user support.

4. System Architecture

The architecture design and infrastructure implementation as an information processing system for big EO data as conceptually outlined above with its main components knowledgebase, factbase, and inference engine raises many technical and practical considerations and challenges. These considerations range from the conceptual means of codifying knowledge, the arrangement and interoperability of many technical components and catering to a diversity of users. On the side of data handling, we must consider data preparation routines, scalability of semantic enrichment, the semantic EO data cube data model(s) and existing EO data cube software. On the side of processing, we must consider the EO data cube software's API, complexity of the inference engine and semantic querying language, the GUI for encoding knowledge, and knowledge engineering in general, including transferability and explainability of semantic models and resulting queries. In terms of infrastructure, we must consider scalability and monitoring processes, security and performance issues. Finally, an expert system requires active users, so we must consider user management, user support, and ideally provide a means for sharing user experiences and building a user community. Figure 2 shows an overview of the architecture and the component orchestration, while the next sub-sections describe the individual components.

4.1. Factbase and Knowledgebase

4.1.1. Data Preparation and Pre-Processing with Semantic Enrichment

The Sen2Cube.at's pre-processing includes the fully automated semantic enrichment of the optical EO imagery. It uses a dockerized version of SIAM that scales depending on the scalable number of images and the resource availability. The pre-processing consists of the following steps:

1. Select and access candidate images. Currently, all available images covering the spatio-temporal target extent of the factbase are selected, regardless of image content (e.g., without filtering for cloud cover);
2. Prepare images for semantic enrichment. Select the necessary spectral bands for each candidate image depending on the sensor, calibrate if not in at least top-of-atmosphere (ToA) reflectance, and transform them into the required format for SIAM. Automatically generate a mask to include only pixels with valid observations in all the necessary bands to guarantee a complete spectral signature and eliminate artefacts and errors at acquisition swath edges;
3. If specified, the input file of necessary spectral bands for SIAM is re-projected using GDAL warp and the bilinear algorithm, and nearest neighbour for the valid-data mask;
4. Semantic enrichment using SIAM. The exact enrichment outputs are derived from the semantic EO data cube's layout [55] (e.g., a total of 33 spectral categories, haze mask, greenness ratio, brightness information layers);
5. Generate metadata files for images and information layers depending on the EO data cube software;
6. Index the new images and information layers into the EO data cube software;
7. Remove intermediate outputs from steps 2–4 depending on configuration.

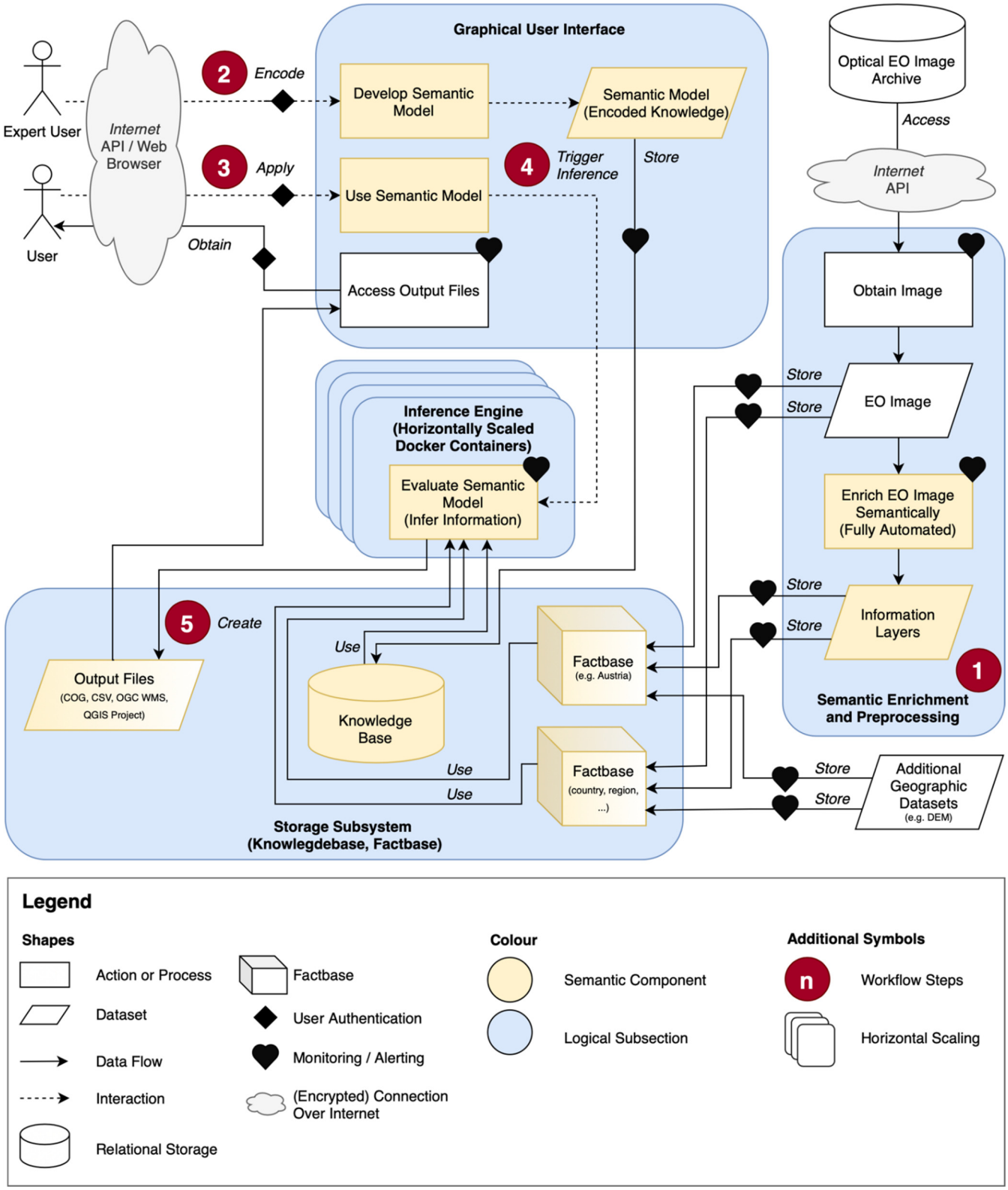

Figure 2. High-level system architecture and component orchestration. The numbers in the figure indicate the sequence of the overall workflow, including semantic enrichment (1), knowledge engineering (2), and semantic querying (3–5).

The output of pre-processing is the semantic enrichment for all EO images in a unified grid and target projection available in the factbase for querying. The indexing of

additional layers, such as a DEM and derived information layers (e.g., retrieval of mountain categories), occurs on a case-by-case basis.

4.1.2. Data Models and Information Management

The data model design aims to appropriately represent the observed geographical phenomena and reduce the costs of computation, storage, and maintenance. The logical representations in Sen2Cube.at are gridded datasets in EO data cubes (factbase) containing data and image-derived information and semantic models in a relational database (knowledgebase). Thus, the thematic information (i.e., interpretation) constitute the thematic dimension of the semantic EO data cube and is either explicitly stored (e.g., semantically enriched images), or implicitly captured in a semantic model. This contrasts with other EO data cubes, where the thematic dimension is not usually augmented with new categorical information and not coupled with a knowledgebase.

The thematic dimension of a semantic EO data cube may vary between instances and requires additional metadata management; therefore, we developed the concept of a layout. A layout describes the semantic EO data cube's structure, value types and basic operations that can be applied (e.g., re-scaling categorical or continuous variables). The layout is human- and machine-readable and used in all system components. For example, the GUI relies on it to automatically enable or disable interfaces, and it defines internal decisions the inference engine makes about product locations or re-sampling strategies [55].

The data models and information management are technology-agnostic, but we selected specific technologies for our implementation. It currently uses the Open Data Cube to instantiate EO data cubes and saves the knowledgebase in PostgreSQL with the semantic models as XML (eXtensible Markup Language) and the layout as JSON (JavaScript Object Notation). The connection with knowledge graphs in the Semantic Web using linked data is conceptually possible, but the observation-based EO data cube would need to be extended with object-based capabilities to fully support such semantic relationships.

4.1.3. Knowledge Engineering Using Semantic Models

The Sen2Cube.at interface allows the development of a semantic model by defining entities and their properties (e.g., color, temporal information) in a visual convergence-of-evidence approach. A semantic model may be application-dependent and incorporate spatial and temporal dimensions because knowledge engineering aims for the spatio-temporal top-down conceptual modelling of real-world entities, in contrast to the generic rule-based data-driven bottom-up initial semantic enrichment. An example that includes the temporal dimension is formulating how to define areas based on where vegetation was lost compared to the previous year.

Our semantic query language is designed to allow users to create semantic models in a dedicated structure. A semantic model in Sen2Cube.at consists of metadata (e.g., name, free-text description, user/owner, timestamp of last update) and two components to allow for generic, automated processing, and comprehensive complex queries (Figure 3):

(1) Semantic concepts: users define semantic concepts, such as entities, by combining several facts as entity properties in a convergence-of-evidence approach (e.g., land cover types). Examples are the spectral categories as color information, texture, topography or spatio-temporal attributes. By definition, and consistent with expert systems, multiple entity properties are connected by a logical AND. Each property contains one or more rules, which can be connected by either AND or OR. Optionally, a rule to define an entity can also be a constraint with additional criteria (e.g., specific time intervals). For example, the land cover type 'surface water' may be defined by attribute values of all available properties (e.g., color, texture, slope, spatio-temporal attributes) while the property 'color' may hold values 'green' or 'blue' and the property 'slope' the value 'flat'. The defined entities may be directly or closely associated with physical entities (e.g., lake, farmland, vegetation). The semantic level is decided by the

user on a case-by-case basis and only limited by the capabilities of available semantic enrichment and additional data (e.g., DEM).

(2) Application: users formulate how to generate analysis results by using one or more defined semantic concepts. Specific, well-defined processing actions can be used and chained to obtain the outputs. Semantic concepts and results can be (re-)used in many outputs in the application domain.

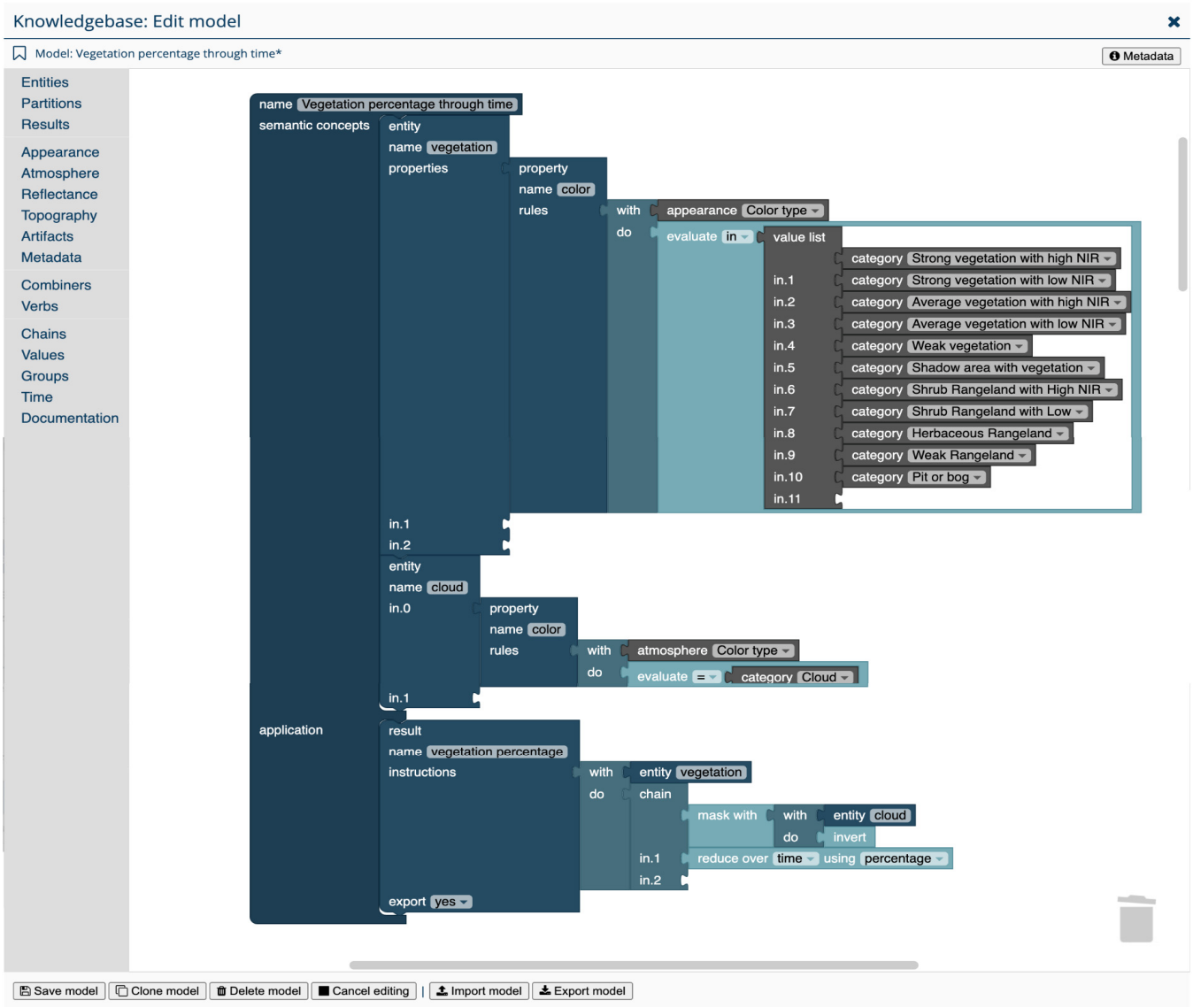

Figure 3. Graphical knowledge engineering approach implemented in our model editor (screenshot). The model editor is based on the Blockly library with our custom blocks for semantic querying. The figure shows the definition of two entities (vegetation and cloud) as semantic concepts based on the properties of combined spectral categories and their analysis through time, which is defined as result in the application part (cloud masked ratio of observed vegetation in any user defined time span and AOI). The semantic model does not require using custom, arbitrary thresholds or image-specific descriptors.

Separation of the definition of semantic concepts definitions and applications (i.e., results) is important to reduce complexity, thus keeping EO analyses tangible. It makes it possible to investigate and improve them independently. For example, it is possible to alter the definition of vegetation while keeping the application part unchanged and vice versa. This is different from typical programming interfaces, which usually start with

loading data and then apply a sequence of technically defined operations that build upon each other. This approach provides a structured yet flexible interface to the user, who can create semantic models to obtain results for queries such as, "count how many times water occurred in each part of my AOI", or, "create a cloud-free composite map for my AOI during a specific time frame".

4.2. Semantic Querying and Inference Engine

The overarching goal of semantic querying is bridging the 4D physical spatio-temporal world and the 2D image domain, which may be extended with the time dimension to extract more information and achieve a higher semantic level. The purpose of the inference engine is to evaluate the semantic models and return the inferred information in data formats that allow interoperability with further processing and/or visualization tasks. The inference engine evaluates a semantic model in a stepwise manner:

1. Prepare semantic model for processing. The inference engine creates its own processing-optimized internal representation of a semantic model. General checks are performed, e.g., whether the query's extent intersects spatio-temporally the factbase's extent;
2. Execute the semantic query. Each building block in a semantic model is associated with a dedicated evaluation function that performs a certain processing task. There are three main types of processing tasks: (1) translating high-level semantic concepts into low-level database queries, (2) using these queries to retrieve actual data values from the factbase, and (3) applying specified data cube operations (e.g., reductions, filters) to the retrieved data;
3. Export outputs. Each result specified in the application part of a semantic model is an n-dimensional array. The number of dimensions (n) depends on the performed reduction and expansion operations. The most common outputs are 1D arrays with a time dimension (i.e., time series such as a vegetation status over a year) and 2D arrays with two spatial dimensions (i.e., map such as green spaces in a city, cloud-free composite). However, other dimension combinations as outputs are possible, including a scalar value (e.g., the maximum number of water pixels in an area), categories (e.g., most occurring entity over time), and dates (e.g., cloud-free dates). For each result, the inference engine writes the content of the returned array to a file. The file type depends on the dimensions of the array. For example, a time series is written to a CSV file, while a map is written to a GeoTIFF file.

The inference engine is developed in-house as a python library that includes a driver to access data and information layers indexed in the ODC. The inference engine uses DataArray objects from the Python library xarray [16] as a data structure for multi-dimensional arrays, and extends the functionalities of xarray to implement evaluation functions for each processing task.

A semantic query consists of the following components:

1. Semantic model: a subset of the knowledgebase that stores users' a-priori knowledge as a set of rules that define semantic concepts (i.e., entities). It also contains a formulation for producing desired results based on these concepts used to infer new information;
2. Area-of-interest: a subset (spatial extent) of the factbase;
3. Time Interval: a subset (temporal extent) of the factbase;
4. Results: one or more outputs, which are defined in the semantic model and generated through inference;
5. Meta-data: information about the inference (e.g., execution times, duration, owner/creator).

Users submit a semantic query with a semantic model either in batch mode, obtaining the results once the query completed, or interactively by developing or changing semantic models and obtaining the results faster. For example, the interactive use-case is helpful in fast prototyping or to create new semantic models in a step-wise manner. To support this, users select the 'preview' mode, which resamples the spatial resolution prior to processing

and significantly reduces the memory consumption and processing time. It is also possible to abort semantic queries, to re-run them or to clone and execute them with adjustments.

4.3. Graphical User Interfaces (GUIs) and Application Programming Interfaces (API)

The web-based GUI (Figure 4) makes it possible to create, organize, and share semantic models in the knowledgebase, apply a semantic model to custom spatio-temporal subsets of a factbase, and view the inference results. The main panels are based on the components of a general architecture of an expert system. The first panel allows users to create or select semantic models in the knowledgebase. The second panel provides interfaces to select the AOI and time interval. AOIs can be drawn directly on the map as a point, line, or polygon in single- or multi-geometries, uploaded as a GeoJSON-encoded file, or added as OGC WFS with GeoJSON encoding. In the third panel, users can start the semantic query. The technical implementation of the GUI is based on the JavaScript framework ember.js (https://emberjs.com/, accessed on 28 October 2021) and extended by various plugins (e.g., leaflet.js for map display).

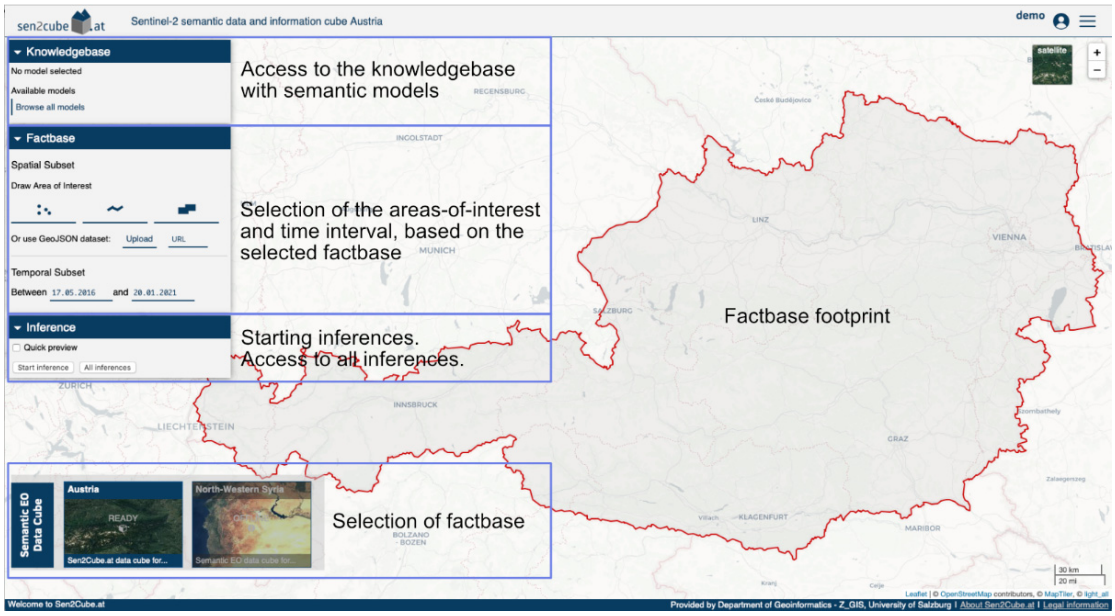

Figure 4. Sen2Cube.at's GUI for semantic querying of big EO data. Here the Austrian factbase is selected, but other instances of factbases can be selected using the same GUI.

The graphical model editor enables knowledge engineering and supports a visual drag-and-drop approach. We chose a graphical approach to facilitate semantic querying in Sen2Cube.at to serve a wide range of potential users (see Figure 1); some users might be domain experts, e.g., in agriculture, but not familiar with programming or EO. Since the initial semantic enrichment of optical EO images is mono-date, application-independent, and only one part of a CV system, the graphical model editor is where users formalize a priori knowledge for their analyses. The graphical model editor is based on the JavaScript library Blockly (https://developers.google.com/blockly/, accessed on 28 October 2021), in which we created new blocks tailored for semantic EO analyses, such as for the semantic models' primitives and grammar.

The GUI provides a simple visualisation for checking inference results (i.e., maps, time series, a single value or list of single values, or calendars). All the results, as well as the AOI, of an inference are packed into an own QGIS project and can therefore be obtained by a variety of interfaces, including download as single files or as a QGIS project, or via service-oriented Open Geospatial Web Service (OWS) access. The interfaces facilitate further processing or map creation and geo-visualisation in external software.

The front end serves as a central hub for multiple instantiations of factbases and users select the factbase that they want for semantic querying [55]. It is a single access points for all factbases and allows the maintenance of a central knowledgebase. The factbases are still separated from each other. A semantic model can be executed on different, but not be across multiple factbases. Distributed system components use encrypted communication over the Internet, which allows deployment of factbases on different cloud infrastructures.

The server-side backend is a Flask REST API with Marshmallow as a data abstraction layer and follows the json:api standard. Therefore, the API is independent of the GUI supporting additional clients (e.g., mobile) or batch-processing (e.g., using command-line interface). API requests need token-based authentication via OAuth2. Secure storage of user credentials and token management are based on Keycloak as an identity server.

Additionally, a command-line interface (CLI) to submit, see, and retrieve inferences. The CLI allows many automatization tasks and the embedding of Sen2Cube.at into existing workflows. Examples include scheduling inferences or batch processing using many or large AOIs. Further, it is also possible to access the factbase using Jupyter notebooks with R and Python kernels as an interactive programming interface, e.g., to use specific packages.

4.4. Performance, Scaling, and Security Considerations

The architecture is designed to facilitate a variety of tasks from diverse users with different requirements. Therefore, the performance criteria do include not only execution speed but also qualitative and quantitative indicators such as reproducibility of results and correctness of evaluation [56].

The architecture allows flexible deployment and fine-grained horizontal scaling (see "Horizontal Scaling" symbol in Figure 2) using Docker containers (e.g., the inference engine and the pre-processing with semantic enrichment). For example, every inference engine process runs in its own container and evaluates only one semantic query at a time, which facilitates concurrent executions. On shared resources, the number of containers is a trade-off between the desired maximum concurrent queries and the maximum resources that a single semantic query can use if queries are executed in parallel.

Special considerations are required during factbase updates, complete resource occupation and factbase crashes. Semantic queries that are executed while new data and information are indexed may return incorrect results if data are read multiple times. Database systems usually provide sophisticated transaction and locking mechanisms to avoid these anomalies. We developed a simple locking mechanism in which factbases can hold different status codes. During normal operation, the factbase status is 'ready', which is changed to 'maintenance' during updates to prevent executing semantic queries. Each semantic query is also assigned a status (e.g., 'created', 'started', 'succeeded', 'failed') and will remain as 'created' if all resources are occupied until some are free again. Thus, the inference engine containers send their occupation to the backend from which a factbase 'busy' or 'ready' status is derived. If no signals are received for longer time, the factbase status is set to 'offline'. The factbase status is visible to users to give an indication why their semantic queries may not be executed.

To prevent overuse by single users, the execution of semantic queries can be limited on a per-user basis, e.g., time limits and resource consumption. This supports different user types ranging from demo users with only a few minutes to users of operational applications with hours or even days of granted execution time. The resource consumption limits are the container settings provided by Docker [57], which can be directly used since every inference engine executes semantic queries in its own container.

The operation of cloud-based big EO data systems with shared resources, fully automated and regular processing, and occasional anomalies in the source data [58], requires a monitoring system that supports alerts. All of this can hamper near-real-time applications. We use Grafana with Loki to investigate log files and Prometheus to monitor resources allowing the investigation of logs of crashes and errors, as well as to track computing resource consumption, detect bottlenecks, identify errors or skipped images during

the automated semantic enrichment, or send alerts if factbase updates with new images unexpectedly stop.

Security considerations of unauthorized data access and manipulation are mitigated by state-of-the-art user management with extended features (e.g., password policy, two-factor authentication, standard TLS/SSL encryption). Further, the semantic queries are executed in read-only containers with read-only mounted directories. The image index database in PostgreSQL is accessed by a database role with limited read privileges on necessary tables. All the user inputs are checked (e.g., intersection with the spatial footprint, start and end dates) and sanitized (e.g., using prepared statements or the slugify Python package).

4.5. Support, Workshops, and Community

The Sen2Cube.at approach involves many user interactions during the creation of semantic models and augmentation of the knowledgebase. Multiple users with different expertise can collaborate, such as domain experts and EO experts. The intention is to reach out to potential users who are new to the EO domain. Therefore, the user perspective is crucial, and we want to support as many users as best as possible by conducting workshops, collecting and integrating user feedback, and providing a detailed online manual (https://manual.sen2cube.at, accessed on 28 October 2021) and an online forum (https://forum.sen2cube.at, accessed on 28 October 2021) where users can obtain information about updates but also ask questions, report problems or discuss use-cases and features.

5. The National Austrian Semantic EO Data Cube Infrastructure

While the presented architecture is not limited to a specific sensor or a specific semantic enrichment, our first instance of this approach is the Austrian semantic EO data cube infrastructure and contains images captured by the Sentinel-2 MSI sensor. At the time of writing, we have semantically enriched and indexed all the available Sentinel-2 L1C products from 14 granules (32TNT, 32TQT, 33TUN, 33TVN, 33TWN, 33UVP, 33UWQ, 32TPT, 33TUM, 33TVM, 33TWM, 33UUP, 33UWP, and 33UXP) acquired from July 2015 to now (ca. 13,000 images). New images are automatically pre-processed, semantically enriched, and added to the factbase on a daily basis. The Sentinel-2 images covering Austria are originally provided in UTM zones 32N and 33N, which we reproject to an equal area projection used by statistical agencies in Europe (EPSG:3035) with targeted aligned pixels, ultimately aligning the 10m output to the EEA reference grid [59]. A 10m spatial resolution DEM covering Austria based on combined provincial airborne laser scanning data (https://www.data.gv.at/katalog/dataset/dgm, accessed on 28 October 2021) is also indexed, as well as derived information (i.e., slope, aspect, surface roughness) that can also be used in semantic models.

Semantic models for different applications are currently being designed and tested in Sen2Cube.at with Sentinel-2 images, but the models are sensor- and geography-independent. This means that the semantic models are largely transferable to factbases based on other optical sensors and areas outside Austria. Queries and applications currently under investigation that may be directly transferred into services include: (a) semantic-content-based image retrieval with AOI-based execution and download-link to original images (e.g., images that have less than 10% cloud in an AOI); (b) the generation of composites using user-defined best-pixel-selection (e.g., cloud-free, greenest); (c) (multi-) parcel-based analysis with maps or time-series as the output (e.g., to detect mowing of grassland or harvest of agricultural fields); (d) national or state-level analyses with maps as output (e.g., indicate potential areas with soil sealing or vegetation loss). Figure 5 shows one output that covers the entire spatial extent of all granules from 1 March 2020 to 30 September 2020 and visualizes the percentage of valid observations (i.e., pixel-based cloud masking using the semantic concepts) where vegetation was observed.

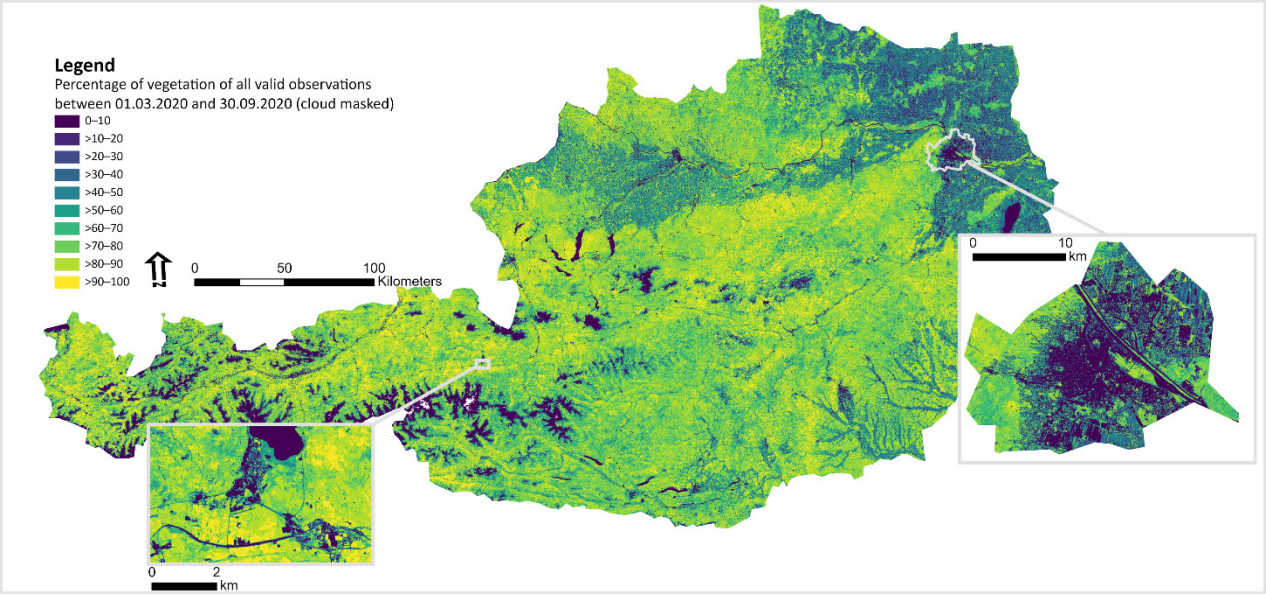

Figure 5. A vegetation percentage over time within the vegetation season (1 March 2020–30 September 2020) based on semantic concepts and with excluded clouds on 10 m ground resolution. Users can define their vegetation concept graphically and semantically, i.e., based on the meaning of the spectral categories and without arbitrary thresholds of vegetation indices as it is required in other non-semantic approaches. It is possible to mask out clouds easily and without relying on per-image statistics or on algorithms with known problems [60]. From the user perspective, this analysis was generated in a web-browser without other hardware than a standard office computer and without programming and can be repeated for other time steps and/or areas by re-using the created model. Inset on the right shows Vienna, inset on the left the southern part of the city Zell am See.

We apply Sen2Cube.at in a variety of applications, including monitoring agricultural practices to detect mowing and harvesting events, indicating soil sealing, calculating the mountain green cover index, and augmenting point-based landslide inventories.

6. Discussion

Our approach to the semantic analysis of big EO data is unique, offers several benefits and goes beyond other current approaches, but it also has some trade-offs and limitations.

It is one of the very few approaches that does not rely on programming in code editors but aims to provide a graphical, coding-free interface in which even custom, user-defined analyses can be conducted without programming skills. Other graphical approaches to analysing big EO data include the OpenEO web editor (https://editor.openeo.org, accessed on 28 October 2021) and Earth Blox (https://www.earthblox.io/, accessed on 28 October 2021). However, analyses in a semantic querying language are closer to the domain language and the design of our language is driven by meaning and concepts (e.g., entities and properties) instead of technical considerations (e.g., image and band names, procedural instructions). For example, initial semantic enrichment with general-purpose spectral categories as initial building blocks avoids arbitrary user-defined thresholds. The structure of our semantic querying language separates entities and their applications, allows for quicker workflow generation with reusable components, and avoids the pitfall of excessive growth of the models, which is a negative experience with many graphical editors. The time to develop workflows and conduct analyses can be significantly lower compared to approaches that start from scratch using a code editor. As a web-application with JSON-API, Sen2Cube.at allows purely browser-based usage, integration into existing workflows, and offers interfaces to other systems.

Our knowledge-based approach contributes to the endeavour of building explainable AI systems for EO data analysis. Within such a system, the initial semantic enrichment is only a first module. We use SIAM to semantically enrich optical EO images, which is a

color naming method. The resulting spectral categories are associated to color properties of (land cover) entities within user-defined semantic models in a convergence-of-evidence approach. Since the knowledgebase is continuously augmented with expert knowledge, the system's AI capability is ever-increasing. Separating the fully automatic, sample-free semantic enrichment from user-defined semantic models avoids the pitfall of some AI expert systems with respect to limited flexibility and complexity. The overall goal is yet to be achieved, since some important entity properties, such as geometric object attributes (e.g., size, shape) and spatio-temporal relationships (e.g., overlap, close-to,) are not yet part of our semantic models and data models. However, our architecture of an AI system for CV of optical EO images is in line with (i.e., a subset of) a full, hierarchical, multi-stage image understanding system. The analysis capabilities of spatial information, which even dominates color information in image analysis [61], can be added. In the big EO data domain, only prototypical work aims to include spatial information [14,44], and more research, e.g., on hybrid data models, is required.

The semantic enrichment increases the volume of the data that needs to be stored, but it reduces the processing time because users do not start from scratch. Categorical variables stored in one band can be represented with fewer bits than the reflectance values of multispectral EO images stored in several bands. Since many semantic queries can be conducted on category values stored in 8Bit raster data instead of multi-band reflectance values stored in 10Bit raster data per band, many analyses are faster with less memory footprint. Typically, cloud storage is cheaper than cloud processing [17]. The more often the data is used the higher the cost-saving effect. To facilitate more implementations, semantic enrichment should happen early in the EO image processing chain (i.e., at the ground segment). Additional data, such as the DEM-derived layers (slope, aspect, roughness), can also be semantically enriched to better facilitate semantic querying based on more meaningful categories. While some categories may already be available [62], their applicability needs to be checked and, if necessary, new categories need to be found or developed.

The approach presented here reaches its limit with very specific applications that require special algorithms or physical models that cannot be formulated in semantic models (e.g., estimating carbon stocks, analysing urban heat islands). However, the approach still may have its merits within a stratified approach that uses land cover maps.

We associate the graphical approach and semantic querying with an increased usability as illustrated in Figure 2 because it is closer to domain language, abstracts many technical implementation details, and allows the sharing of semantic models between expert and non-expert users. Experienced users with programming skills may prefer a programming interface; there are lots of examples of this, including the SITS R package [32] or OpenEO [63]. A programmatic approach to creating semantic models is currently under development, although it focuses on the semantic analyses and therefore will remain conceptually different from these developments.

New factbases outside Austria, including automated semantic enrichment, have already been created for some smaller areas (e.g., covering parts of Syria [64] and Afghanistan) and other sensors (e.g., AVHRR-3). An automated approach to creating an on-demand factbase in a cloud environment is also currently under development, similar to the approach by [65].

Since our Sen2Cube.at architecture is cloud-based, it remains affected by all the typical limitations of cloud-based systems, including the fact that users are not in full control of certain steps, as well as the data, or are affected by internet connectivity problems [7].

7. Conclusions and Outlook

We developed Sen2Cube.at, a scalable semantic EO data cube architecture, as a step towards the production of an explainable AI-based CV system for big EO data, implementing it prototypically for Sentinel-2 images covering Austria. This is the first implementation of a semantic EO data cube on a national level worldwide. It includes the automated genera-

tion of application-generic information layers (i.e., semantic enrichment) and a semantic querying language within a GUI, which allows users to query the semantic EO data cube in a graphical way and as close as possible to their own domain language.

This semantics-enabled approach differs from existing multi-temporal big EO data analysis approaches because it is general-purpose and based on semantic querying using a priori knowledge, meaning it is an explainable AI approach. Semantic querying works differently to approaches based on procedural or declarative code. In contrast to non-semantic approaches, SCBIR based on any semantic concepts, e.g., cloud-free custom AOIs, maximum surface water (flooded areas), and strongest vegetation, is possible. With respect to image analysis, we showed the benefits using a simple example of inferring vegetation occurrence. The semantic model for this semantic query does not require users to know internal descriptors (e.g., band names) and execution steps or to set arbitrary, image- and location-dependent vegetation index thresholds without a meaning (e.g., 0.3 or 0.4). It allows the exclusion of noise, e.g., image-independent cloud masks based on the spectral categories and defined semantic concepts, as a criterion in the same way. In our semantic querying, the target spatio-temporal entities are defined as general, data-independent, transferable knowledge using a convergence-of-evidence approach, in which color is just one property or piece of evidence. The hierarchical CV system does not associate the spectral signature of reflectance values directly to land cover types; rather, it considers the spectral signatures as a color property of spatio-temporal entities.

In addition to the technical and scientific differences between approaches, usability plays a vital role for end-users depending on user type [5]. The semantic models divide the complexity of an EO analysis into tangible pieces, formulated in a more "natural", meaningful way and facilitated by a GUI. New information can be inferred by an automated translation of target entity definitions into queries against a factbase. Thus, it allows even complex querying and offers the potential to reach and engage new and non-EO expert users, contributing to an increased uptake of free and open big EO data archives. Our own experiences with users and their feedback so far have been positive; however, an objective evaluation of usability and the comparison with different (non-semantic) approaches, including various code editors (e.g., openEO, GEE, ODC Jupyter notebooks) and other visual approaches is not easy to obtain. Since the number of available platforms for analyzing big EO data are increasing, the end users as well as the entire community would benefit from a dedicated usability study that compare various approaches and platforms, considering different types of users and their requirements.

In many applications, Sen2Cube.at will only be part of a larger workflow with additional methods and tools. For example, either the input AOI is the result of a GIS operation or the output data is further processed or combined with other data. This is similar to other EO data cube implementations. Therefore, we recognize the importance of standardized input and output interfaces that allow the seamless integration of Sen2Cube.at into (existing) workflows. Examples of these interfaces are OGC WFS for AOI inputs and obtaining results in machine-readable formats or OGC WMS. It is also possible to switch to a Jupyter notebook and directly use the results in further R or Python processing.

Similar to experiences gathered by the Open Data Cube Initiative [17] and Digital Earth Africa [66], we emphasize that merely providing access to the software is not sufficient to meet users' needs, particularly with a new and different approach. The initial development of such an architecture can be achieved by research projects, but the transition into a sustainable operation as infrastructure requires different tasks. These tasks include continuous software maintenance, bug-fixing and improvements, monitoring, data updates, as well as up-to-date documentation, a manual, information and training material, a community forum, and workshops. The sustainable operation shifts the focus and type of required effort and affects access to funding, knowledge, and the skills of the people involved, as well as the stakeholders.

Future technical developments will include general upscaling, the operationalization of the approach, and the on-demand instantiation of new factbases for any area of the

Earth. This requires the further development and implementation of the layout concept. Supporting different use-cases and application scenarios will be achieved by adding more data and information (e.g., different sensors), performing data fusion on the semantic level (e.g., queries across multiple factbases), and developing additional interfaces (e.g., GIS, mobile applications). Further, a more sophisticated mechanism for scheduling semantic queries, including notifications, together with an AOI database will improve monitoring and reporting applications. Improvements in the performance include parallelisation of the inference engine, using cloud-optimized data formats, such as Cloud-optimized-GeoTIFFs that support pyramids and tiling, and STAC (Spatio-Temporal Asset Catalog).

Scientific research will include the investigation of the automated reporting of machine-readable provenance graphs of inference results, which goes beyond current metadata. Such a provenance graph typically includes the source datasets and processing steps [67,68]. While this is supposed to increase trust in results and address the veracity of big data in general, a research gap remains the development of methods and tools that evaluate the provenance graph on the client side. Extensions of the data model include object-based approaches that go beyond parcel-based queries to allow using object attributes and spatial relationships in the semantic models. This could facilitate the linking to knowledge graphs, e.g., with GeoSPARQL queries. Further, we will investigate the applicability of the Sen2Cube.at system in more applications, including those that are related to SDG indicators, such as the mountain green cover index (15.4.2) or the soil sealing index (15.4.1).

Author Contributions: Conceptualization, M.S., A.B., H.A. and D.T.; methodology, all; software, M.S., A.B., L.v.d.M. and H.A.; writing—original draft preparation, M.S.; writing—review and editing, all; visualization, M.S. and D.T.; project administration, D.T.; funding acquisition, D.T., M.S. and H.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Austrian Science Fund (FWF) through the GIScience Doctoral College (DK W 1237-N23) and by the BMK/Austrian Research Promotion Agency (FFG) under the Austrian Space Application Programme (ASAP) within the project Sen2Cube.at (project no.: 866016).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The Austrian semantic EO data cube infrastructure can be accessed under https://sen2cube.at (accessed 25 November 2021).

Acknowledgments: We thank Steffen Reichel and Christian Werner for code contributions to the implementation.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Dhu, T.; Guiliani, G.; Juárez, J.; Kavvada, A.; Killough, B.; Merodio, P.; Minchin, S.; Ramage, S. National Open Data Cubes and Their Contribution to Country-Level Development Policies and Practices. *Data* **2019**, *4*, 144. [CrossRef]
2. Lewis, A.; Lacey, J.; Mecklenburg, S.; Ross, J.; Siqueira, A.; Killough, B.; Szantoi, Z.; Tadono, T.; Rosenavist, A.; Goryl, P.; et al. CEOS Analysis Ready Data for Land (CARD4L) Overview. In Proceedings of the IGARSS 2018—2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain, 22–27 July 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 7407–7410.
3. Mueller, N.; Lewis, A.; Roberts, D.; Ring, S.; Melrose, R.; Sixsmith, J.; Lymburner, L.; McIntyre, A.; Tan, P.; Curnow, S.; et al. Water Observations from Space: Mapping Surface Water from 25 Years of Landsat Imagery across Australia. *Remote Sens. Environ.* **2016**, *174*, 341–352. [CrossRef]
4. Baraldi, A.; Boschetti, L. Operational Automatic Remote Sensing Image Understanding Systems: Beyond Geographic Object-Based and Object-Oriented Image Analysis (GEOBIA/GEOOIA). Part 1: Introduction. *Remote Sens.* **2012**, *4*, 2694–2735. [CrossRef]
5. Wagemann, J.; Siemen, S.; Seeger, B.; Bendix, J. Users of Open Big Earth Data—An Analysis of the Current State. *Comput. Geosci.* **2021**, *157*, 104916. [CrossRef]
6. Kavvada, A.; Metternicht, G.; Kerblat, F.; Mudau, N.; Haldorson, M.; Laldaparsad, S.; Friedl, L.; Held, A.; Chuvieco, E. Towards Delivering on the Sustainable Development Goals Using Earth Observations. *Remote Sens. Environ.* **2020**, *247*, 111930. [CrossRef]

7. Sudmanns, M.; Tiede, D.; Lang, S.; Bergstedt, H.; Trost, G.; Augustin, H.; Baraldi, A.; Blaschke, T. Big Earth Data: Disruptive Changes in Earth Observation Data Management and Analysis? *Int. J. Digit. Earth* **2019**, *13*, 832–850. [CrossRef] [PubMed]
8. Kluyver, T.; Ragan-Kelley, B.; Pérez, F.; Granger, B.E.; Bussonnier, M.; Frederic, J.; Kelley, K.; Hamrick, J.B.; Grout, J.; Corlay, S.; et al. Jupyter Notebooks—A Publishing Format for Reproducible Computational Workflows. In *Positioning and Power in Academic, Proceedings of the 20th International Conference on Electronic Publishing, Göttingen, Germany, 7–9 June 2016*; Loizides, F., Schmidt, B., Eds.; IOS Press: Berlin, Germany; Washington, DC, USA, 2016; pp. 87–90.
9. Baumann, P. The OGC Web Coverage Processing Service (WCPS) Standard. *Geoinformatica* **2010**, *14*, 447–479. [CrossRef]
10. Gorelick, N.; Hancher, M.; Dixon, M.; Ilyushchenko, S.; Thau, D.; Moore, R. Google Earth Engine: Planetary-Scale Geospatial Analysis for Everyone. *Remote Sens. Environ.* **2017**, *202*, 18–27. [CrossRef]
11. Valero, S.; Morin, D.; Inglada, J.; Sepulcre, G.; Arias, M.; Hagolle, O.; Dedieu, G.; Bontemps, S.; Defourny, P.; Koetz, B. Production of a Dynamic Cropland Mask by Processing Remote Sensing Image Series at High Temporal and Spatial Resolutions. *Remote Sens.* **2016**, *8*, 55. [CrossRef]
12. Gascon, F. Sentinel-2 for Agricultural Monitoring. In Proceedings of the IGARSS 2018—2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain, 22–27 July 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 8166–8168.
13. Arvor, D.; Betbeder, J.; Daher, F.R.G.; Blossier, T.; Le Roux, R.; Corgne, S.; Corpetti, T.; de Freitas Silgueiro, V.; Silva Junior, C.A. da Towards User-Adaptive Remote Sensing: Knowledge-Driven Automatic Classification of Sentinel-2 Time Series. *Remote Sens. Environ.* **2021**, *264*, 112615. [CrossRef]
14. Baraldi, A.; Tiede, D. AutoCloud+, a "Universal" Physical and Statistical Model-Based 2D Spatial Topology-Preserving Software for Cloud/Cloud–Shadow Detection in Multi-Sensor Single-Date Earth Observation Multi-Spectral Imagery—Part 1: Systematic ESA EO Level 2 Product Generati. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 457. [CrossRef]
15. Augustin, H.; Sudmanns, M.; Tiede, D.; Lang, S.; Baraldi, A. Semantic Earth Observation Data Cubes. *Data* **2019**, *4*, 102. [CrossRef]
16. Hoyer, S.; Hamman, J.J. Xarray: N-D Labeled Arrays and Datasets in Python. *J. Open Res. Softw.* **2017**, *5*, 10. [CrossRef]
17. Killough, B. Overview of the Open Data Cube Initiative. In Proceedings of the IGARSS 2018—2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain, 22–27 July 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 8629–8632.
18. Dhu, T.; Dunn, B.; Lewis, B.; Lymburner, L.; Mueller, N.; Telfer, E.; Lewis, A.; McIntyre, A.; Minchin, S.; Phillips, C. Digital Earth Australia—Unlocking New Value from Earth Observation Data. *Big Earth Data* **2017**, *1*, 64–74. [CrossRef]
19. Giuliani, G.; Chatenoux, B.; De Bono, A.; Rodila, D.; Richard, J.-P.; Allenbach, K.; Dao, H.; Peduzzi, P. Building an Earth Observations Data Cube: Lessons Learned from the Swiss Data Cube (SDC) on Generating Analysis Ready Data (ARD). *Big Earth Data* **2017**, *1*, 100–117. [CrossRef]
20. Giuliani, G.; Chatenoux, B.; Honeck, E.; Richard, J.-P. Towards Sentinel-2 Analysis Ready Data: A Swiss Data Cube Perspective. In Proceedings of the IGARSS 2018—2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain, 22–27 July 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 8659–8662.
21. Ariza-Porras, C.; Bravo, G.; Villamizar, M.; Moreno, A.; Castro, H.; Galindo, G.; Cabera, E.; Valbuena, S.; Lozano, P. CDCol: A Geoscience Data Cube that Meets Colombian Needs. In *Advances in Computing. CCC 2017. Communications in Computer and Information Science*; Solano, A., Ordoñez, H., Eds.; Springer: Cham, Switzerland, 2017; pp. 87–99.
22. Quang, N.H.; Tuan, V.A.; Hao, N.T.P.; Hang, L.T.T.; Hung, N.M.; Anh, V.L.; Phuong, L.T.M.; Carrie, R. Synthetic Aperture Radar and Optical Remote Sensing Image Fusion for Flood Monitoring in the Vietnam Lower Mekong Basin: A Prototype Application for the Vietnam Open Data Cube. *Eur. J. Remote Sens.* **2019**, *52*, 599–612. [CrossRef]
23. Asmaryan, S.; Muradyan, V.; Tepanosyan, G.; Hovsepyan, A.; Saghatelyan, A.; Astsatryan, H.; Grigoryan, H.; Abrahamyan, R.; Guigoz, Y.; Giuliani, G. Paving the Way towards an Armenian Data Cube. *Data* **2019**, *4*, 117. [CrossRef]
24. Maso, J.; Zabala, A.; Serral, I.; Pons, X. A Portal Offering Standard Visualization and Analysis on Top of an Open Data Cube for Sub-National Regions: The Catalan Data Cube Example. *Data* **2019**, *4*, 96. [CrossRef]
25. Euro Data Cube Consortium Euro Data Cube. Available online: https://eurodatacube.com (accessed on 28 October 2021).
26. Mahecha, M.D.; Gans, F.; Brandt, G.; Christiansen, R.; Cornell, S.E.; Fomferra, N.; Kraemer, G.; Peters, J.; Bodesheim, P.; Camps-Valls, G.; et al. Earth System Data Cubes Unravel Global Multivariate Dynamics. *Earth Syst. Dyn.* **2020**, *11*, 201–234. [CrossRef]
27. Baumann, P.; Dehmel, A.; Furtado, P.; Ritsch, R.; Widmann, N. The Multidimensional Database System RasDaMan. In *Acm Sigmod Record*; ACM: New York, NY, USA, 1998; Volume 27, pp. 575–577.
28. Baumann, P.; Mazzetti, P.; Ungar, J.; Barbera, R.; Barboni, D.; Beccati, A.; Bigagli, L.; Boldrini, E.; Bruno, R.; Calanducci, A.; et al. Big Data Analytics for Earth Sciences: The EarthServer Approach. *Int. J. Digit. Earth* **2016**, *9*, 1–27. [CrossRef]
29. Storch, T.; Reck, C.; Holzwarth, S.; Wiegers, B.; Mandery, N.; Raape, U.; Strobl, C.; Volkmann, R.; Böttcher, M.; Hirner, A.; et al. Insights into CODE-DE—Germany's Copernicus Data and Exploitation Platform. *Big Earth Data* **2019**, *3*, 338–361. [CrossRef]
30. Soille, P.; Burger, A.; De Marchi, D.; Kempeneers, P.; Rodriguez, D.; Syrris, V.; Vasilev, V. A Versatile Data-Intensive Computing Platform for Information Retrieval from Big Geospatial Data. *Futur. Gener. Comput. Syst.* **2018**, *81*, 30–40. [CrossRef]
31. Guo, H. Big Earth Data: A New Frontier in Earth and Information Sciences. *Big Earth Data* **2017**, *1*, 4–20. [CrossRef]
32. Simoes, R.; Camara, G.; Queiroz, G.; Souza, F.; Andrade, P.R.; Santos, L.; Carvalho, A.; Ferreira, K. Satellite Image Time Series Analysis for Big Earth Observation Data. *Remote Sens.* **2021**, *13*, 2428. [CrossRef]

33. Baumann, P.; Misev, D.; Merticariu, V.; Huu, B.P.; Bell, B. DataCubes: A Technology Survey. In Proceedings of the IGARSS 2018—2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain, 22–27 July 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 430–433.
34. Gomes, V.C.F.; Queiroz, G.R.; Ferreira, K.R. An Overview of Platforms for Big Earth Observation Data Management and Analysis. *Remote Sens.* **2020**, *12*, 1253. [CrossRef]
35. Datcu, M.; Daschiel, H.; Pelizzari, A.; Quartulli, M.; Galoppo, A.; Colapicchioni, A.; Pastori, M.; Seidel, K.; Marchetti, P.G.; D'Elia, S. Information Mining in Remote Sensing Image Archives: System Concepts. *IEEE Trans. Geosci. Remote Sens.* **2003**, *41*, 2923–2936. [CrossRef]
36. Li, Y.; Bretschneider, T. Semantics-Based Satellite Image Retrieval Using Low-Level Features. In Proceedings of the 2004 IGARSS—2004 IEEE International Geoscience and Remote Sensing Symposium, Anchorage, AK, USA, 20–24 September 2004; IEEE: Piscataway, NJ, USA, 2004; Volume 7, pp. 4406–4409.
37. Dumitru, C.O.; Cui, S.; Schwarz, G.; Datcu, M. Information Content of Very-High-Resolution SAR Images: Semantics, Geospatial Context, and Ontologies. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2015**, *8*, 1635–1650. [CrossRef]
38. Alirezaie, M.; Kiselev, A.; Längkvist, M.; Klügl, F.; Loutfi, A. An Ontology-Based Reasoning Framework for Querying Satellite Images for Disaster Monitoring. *Sensors* **2017**, *17*, 2545. [CrossRef]
39. Tran, B.-H.; Aussenac-Gilles, N.; Comparot, C.; Trojahn, C. Semantic Integration of Raster Data for Earth Observation: An RDF Dataset of Territorial Unit Versions with Their Land Cover. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 503. [CrossRef]
40. Woodcock, C.E.; Loveland, T.R.; Herold, M.; Bauer, M.E. Transitioning from Change Detection to Monitoring with Remote Sensing: A Paradigm Shift. *Remote Sens. Environ.* **2020**, *238*, 111558. [CrossRef]
41. Bullock, E.L.; Woodcock, C.E.; Holden, C.E. Improved Change Monitoring Using an Ensemble of Time Series Algorithms. *Remote Sens. Environ.* **2020**, *238*, 111165. [CrossRef]
42. Boschetti, M.; Stroppiana, D.; Brivio, P.A.; Bocchi, S. Multi-Year Monitoring of Rice Crop Phenology through Time Series Analysis of MODIS Images. *Int. J. Remote Sens.* **2009**, *30*, 4643–4662. [CrossRef]
43. Griffiths, P.; Nendel, C.; Pickert, J.; Hostert, P. Towards National-Scale Characterization of Grassland Use Intensity from Integrated Sentinel-2 and Landsat Time Series. *Remote Sens. Environ.* **2020**, *238*, 111124. [CrossRef]
44. Tiede, D.; Baraldi, A.; Sudmanns, M.; Belgiu, M.; Lang, S. Architecture and Prototypical Implementation of a Semantic Querying System for Big Earth Observation Image Bases. *Eur. J. Remote Sens.* **2017**, *50*, 452–463. [CrossRef] [PubMed]
45. Maciel, A.M.; Camara, G.; Vinhas, L.; Picoli, M.C.A.; Begotti, R.A.; de Assis, L.F.F.G. A Spatiotemporal Calculus for Reasoning about Land-Use Trajectories A Spatiotemporal Calculus for Reasoning about Land-Use Trajectories. *Int. J. Geogr. Inf. Sci.* **2018**, *33*, 176–192. [CrossRef]
46. Nativi, S.; Mazzetti, P.; Craglia, M. A View-Based Model of Data-Cube to Support Big Earth Data Systems Interoperability. *Big Earth Data* **2017**, *1*, 75–99. [CrossRef]
47. Lewis, A.; Lymburner, L.; Purss, M.B.J.; Brooke, B.; Evans, B.; Ip, A.; Dekker, A.G.; Irons, J.R.; Minchin, S.; Mueller, N.; et al. Rapid, High-Resolution Detection of Environmental Change over Continental Scales from Satellite Data—The Earth Observation Data Cube. *Int. J. Digit. Earth* **2016**, *9*, 106–111. [CrossRef]
48. Strobl, P.; Baumann, P.; Lewis, A.; Szantoi, Z.; Killough, B.; Purss, M.; Craglia, M.; Nativi, S.; Held, A.; Dhu, T. The six faces of the data cube. In *Proceedings of the 2017 Conference on Big Data from Space*; Soille, P., Marchetti, P., Eds.; Publications Office of the European Union: Luxembourg, 2017; pp. 32–35, ISBN 978-92-79-73527-1.
49. Baraldi, A.; Humber, M.L.; Tiede, D.; Lang, S. GEO-CEOS Stage 4 Validation of the Satellite Image Automatic Mapper Lightweight Computer Program for ESA Earth Observation Level 2 Product Generation—Part 1: Theory. *Cogent Geosci.* **2018**, *4*, 1467357. [CrossRef]
50. Baraldi, A.; Humber, M.L.; Tiede, D.; Lang, S. GEO-CEOS Stage 4 Validation of the Satellite Image Automatic Mapper Lightweight Computer Program for ESA Earth Observation Level 2 Product Generation—Part 2: Validation. *Cogent Geosci.* **2018**, *4*, 1467254. [CrossRef]
51. Baraldi, A.; Humber, M.; Boschetti, L. Quality Assessment of Pre-Classification Maps Generated from Spaceborne/Airborne Multi-Spectral Images by the Satellite Image Automatic Mapper^{TM} and Atmospheric/Topographic Correction^{TM}-Spectral Classification Software Products: Part 2—Experimental Result. *Remote Sens.* **2013**, *5*, 5209–5264. [CrossRef]
52. Baraldi, A.; Durieux, L.; Simonetti, D.; Conchedda, G.; Holecz, F.; Blonda, P. Automatic Spectral-Rule-Based Preliminary Classification of Radiometrically Calibrated SPOT-4/-5/IRS, AVHRR/MSG, AATSR, IKONOS/QuickBird/OrbView/GeoEye, and DMC/SPOT-1/-2 Imagery—Part I: System Design and Implementation. *IEEE Trans. Geosci. Remote Sens.* **2010**, *48*, 1299–1325. [CrossRef]
53. Baraldi, A.; Durieux, L.; Simonetti, D.; Conchedda, G.; Holecz, F.; Blonda, P. Automatic Spectral Rule-Based Preliminary Classification of Radiometrically Calibrated SPOT-4/-5/IRS, AVHRR/MSG, AATSR, IKONOS/QuickBird/OrbView/GeoEye, and DMC/SPOT-1/-2 Imagery—Part II: Classification Accuracy Assessment. *IEEE Trans. Geosci. Remote Sens.* **2010**, *48*, 1326–1354. [CrossRef]
54. Laurini, R.; Thompson, D. *Fundamentals of Spatial Information Systems*; Academic Press: London, UK, 1992; ISBN 9780124383807.
55. Sudmanns, M.; Augustin, H.; van der Meer, L.; Werner, C.; Baraldi, A.; Tiede, D. One GUI to Rule Them All: Accessing Multiple Semantic EO Data Cubes in One Graphical User Interface. *GI_Forum* **2021**, *1*, 53–59. [CrossRef]

56. Baraldi, A.; Tiede, D.; Sudmanns, M.; Belgiu, M.; Lang, S. Systematic ESA EO Level 2 Product Generation as Pre-Condition to Semantic Content-Based Image Retrieval and Information/Knowledge Discovery in EO Image Databases. *Proc. BiDS* **2017**, *17*, 17–20.
57. Docker Runtime Options with Memory, CPUs, and GPUs. Available online: https://github.com/docker/docker.github.io/blob/df661019a6f56d9ce3fc2053cad0b6f05802f9e4/config/containers/resource_constraints.md (accessed on 28 October 2021).
58. ESA. Sentinel-2 L1C Data Quality Report Issue 65 (July 2021). 2021. Available online: https://sentinels.copernicus.eu/documents/247904/685211/Sentinel-2_L1C_Data_Quality_Report.pdf/6ad66f15-48ca-4e65-b304-59ef00b7f0e0 (accessed on 28 October 2021).
59. Pfeifer, H. About the EEA Reference Grid. 2021. Available online: https://www.eea.europa.eu/data-and-maps/data/eea-reference-grids-2/about-the-eea-reference-grid/eea_reference_grid_v1.pdf (accessed on 28 October 2021).
60. Tiede, D.; Sudmanns, M.; Augustin, H.; Baraldi, A. Investigating ESA Sentinel-2 Products' Systematic Cloud Cover Overestimation in Very High Altitude Areas. *Remote Sens. Environ.* **2021**, *252*, 112163. [CrossRef]
61. Matsuyama, T.; Hwang, V.S.-S. *SIGMA—A Knowledge-Based Aerial Image Understanding System*; Springer: Boston, MA, USA, 1990; ISBN 978-1-4899-0869-8.
62. Kapos, V.; Rhind, J.; Edwards, M.; Price, M.F.; Ravilious, C. Developing a map of the world's mountain forests. In *Forests in Sustainable Mountain Development: A State of Knowledge Report for 2000. Task Force on Forests in Sustainable Mountain Development*; CABI: Wallingford, UK, 2000; pp. 4–19.
63. Schramm, M.; Pebesma, E.; Milenković, M.; Foresta, L.; Dries, J.; Jacob, A.; Wagner, W.; Mohr, M.; Neteler, M.; Kadunc, M.; et al. The OpenEO API–Harmonising the Use of Earth Observation Cloud Services Using Virtual Data Cube Functionalities. *Remote Sens.* **2021**, *13*, 1125. [CrossRef]
64. Augustin, H.; Sudmanns, M.; Tiede, D.; Baraldi, A. A Semantic Earth Observation Data Cube for Monitoring Environmental Changes during the Syrian Conflict. *GI_Forum* **2018**, *1*, 214–227. [CrossRef]
65. Giuliani, G.; Chatenoux, B.; Piller, T.; Moser, F.; Lacroix, P. Data Cube on Demand (DCoD): Generating an Earth Observation Data Cube Anywhere in the World. *Int. J. Appl. Earth Obs. Geoinf.* **2020**, *87*, 102035. [CrossRef]
66. Digital Earth Africa. Digital Earth Africa Phase 1 Summary. 2019. Available online: https://www.digitalearthafrica.org/sites/default/files/downloads/201905_Digital_Earth_Africa_phase1.pdf (accessed on 28 October 2021).
67. Closa, G.; Masó, J.; Proß, B.; Pons, X. W3C PROV to Describe Provenance at the Dataset, Feature and Attribute Levels in a Distributed Environment. *Comput. Environ. Urban Syst.* **2017**, *64*, 103–117. [CrossRef]
68. Figgemeier, H.; Henzen, C.; Rümmler, A. A Geo-Dashboard Concept for the Interactively Linked Visualization of Provenance and Data Quality for Geospatial Datasets. *Agil. GISci. Ser.* **2021**, *2*, 1–8. [CrossRef]