Semantic Data Cubes Utilising Free and Open-Access EO-Data to Generate Spatially-Explicit Evidence for Environmental Monitoring: Applied Use-Case in Syria Based on Sentinel-2 Data

HANNAH AUGUSTIN[1], MARTIN SUDMANNS[1], DIRK TIEDE[1] & ANDREA BARALDI[2]

Abstract: Collections of free and open-access Earth observation (EO) data with global coverage are growing with increasingly higher spatial resolutions and temporal frequency. They are one of the few globally consistent data sources available for generating information in support of international initiatives. However, these data require automated workflows for handling, processing and analysis, including methods to convert data into valid information.

A proof-of-concept implementation of a semantic EO data cube is presented using Open Data Cube technology to generate ad-hoc reproducible, scalable, repeatable and spatially-explicit information as indicators to support monitoring international environmental initiatives. Sentinel-2 data for the study area in north-western Syria (30,000km²) are incorporated daily, including automatically generated generic semantic enrichment. As of December 2018, this encompasses over 800 scenes (~480GB unprocessed Sentinel-2 data). A semantic query resulting in a normalised vegetation index of occurrence over 3 years for the entire study area is demonstrated as a proof-of-concept example.

1 Context

Technological advances have driven many changes in Earth observation (EO) from space, including new, innovative sensors and ways to handle, store and process rapidly growing data archives. In 1972, the launch of the Landsat programme began what is now the longest record of Earth's status and dynamics from space. Opening the archive to public access in 2008 set the stage for free and open EO data (WULDER et al. 2012). The Sentinel satellites from the EU's EO programme, Copernicus, have provided data since 2014. This has increased the spatial resolution and temporal frequency of freely and openly available EO data from a variety of sensors, including radar and multi-spectral instruments. Often referred to as big Earth data, the challenges this type of data poses researchers are not merely technological, in terms of data storage, access and processing, but rather in developing methods to distill *information* from this wealth of data.

Free and open, high-resolution EO data are ideal sources of evidence for generating meaningful information products to support decision-makers in an international context. They provide consistent global coverage, independent of political or other human-imposed borders, offering potential for large-scale, multi-temporal and persistent monitoring and analysis, especially with continued data acquisition for years to come (DRUSCH et al. 2012). The data are constantly gathered without requiring tasking or direct acquisition costs for those utilising them, and their high

[1] Department of Geoinformatics -- Z_GIS, University of Salzburg, Schillerstr. 30, A-5020 Salzburg, E-Mail: [hannah.augustin, martin.sudmanns, dirk.tiede]@sbg.ac.at

[2] Italian Space Agency (ASI), Via del Politecnico, I-00133 Rome RM, E-Mail: andrea6311@gmail.com

temporal frequency could improve timeliness of actionable information for global initiatives given automated information extraction processes (e.g. automated-prior-knowledge based or machine learning classification procedures).
Semantic enrichment (i.e. generating meaningful information) is necessary for turning data into understandable information products. Optical EO data (e.g. Sentinel-2) cannot directly measure most objects, processes or events on Earth because digital numbers are not a standard unit and many different surfaces can be represented by similar values. Indicator extraction is one way to translate this data into meaningful information. Automatic spectral categorisation (i.e. preliminary classification) is one existing transferable method for initial, generic semantic enrichment that can be used to generate indicators (BARALDI et al. 2010).
Automated workflows are necessary for handling the Sentinel-2 mission's expected 3.4TB of daily data *volume* (EUROPEAN SPACE AGENCY 2017). The data have a relatively high *velocity* due to global coverage, on average every five days at the equator, and quite a *variety* in terms of consistency and quality levels (e.g. cloud coverage differs depending on the spatio-temporal location) (SOILLE et al. 2018).
The work presented here is an example of an automated, reproducible framework for handling and analysing big EO data. It demonstrates the benefits of automated, knowledge-based, generic semantic enrichment as the basis for multi-temporal, spatial, semantic queries to produce diverse, transferable, semi-automated indicators for a variety of domains, including environmental monitoring.

2 Applied Use-Case

The north-western Syrian border to Turkey was chosen as the use-case location based on existing EO-based research (TIEDE et al. 2014), low average cloud cover in the Sentinel-2 archive, and the currently on-going conflict, which makes other methods of data acquisition challenging or impossible. The area is comprised of three Sentinel-2 granules covering over 30,000km² (latitudes 36.01°-37.05°N; longitudes 35.67°-39.11°E). Data included is from 28 June 2015 until the time of writing (19 December 2018; see Fig. 1) and incorporates any new data available in the Copernicus Open Access Data Hub on a daily basis. Two relative orbits cover this area resulting in temporally denser data availability where they overlap (Fig. 2).

2.1 Proof-of-Concept Implementation

The workflow implemented here focuses on automation and big data. It encompasses automated downloads from the Copernicus Open Access Hub, re-formatting Sentinel-2 data for processing, preliminarily classification with the Satellite Image Automatic Mapper (SIAM™) (i.e. generating multiple information layers), indexing Sentinel-2 scenes and information layers into an implementation of the ODC and ingesting information layers (Fig. 3). This process runs automatically every day for each of the three study area Sentinel-2 granules (37SBA, 37SCA, 37SDA). The result is daily incorporation of the most recently available Sentinel-2 data ready for analysis that can include semantic queries. Queries are facilitated using Jupyter notebooks by accessing the ingested information layers produced by SIAM™ via the ODC's python API. At the time or

writing, over 800 scenes (ca. 480GB unprocessed Sentinel-2 data) and their information layers are ingested and can be queried.

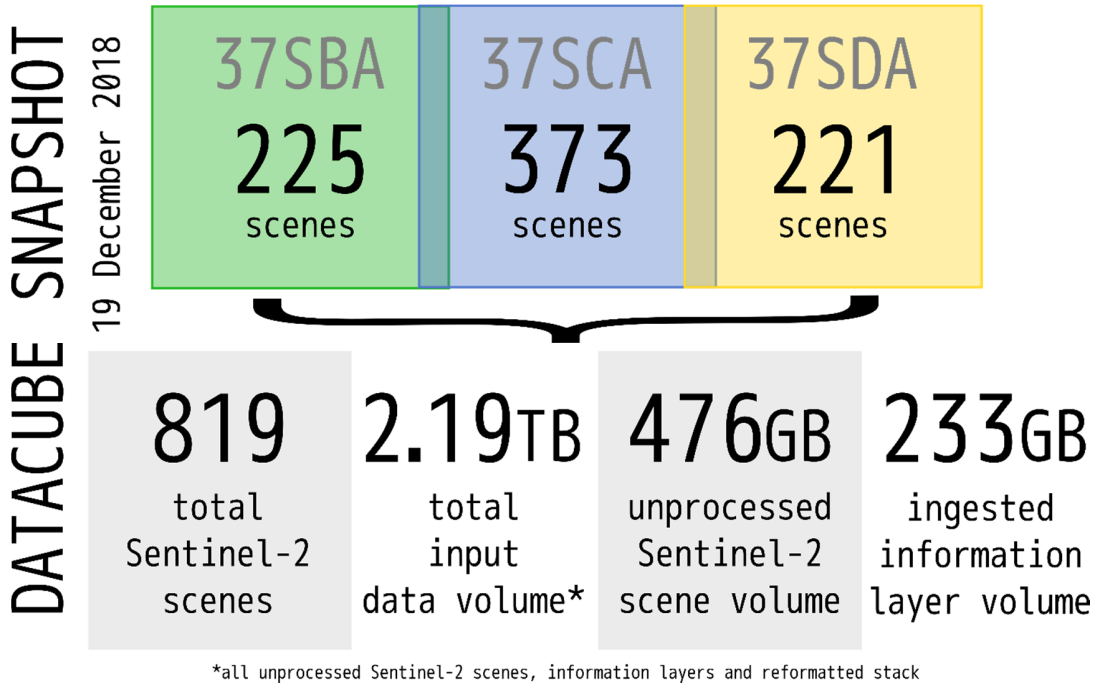

Fig. 1: Snapshot of data included in the implementation as of 19 December 2018

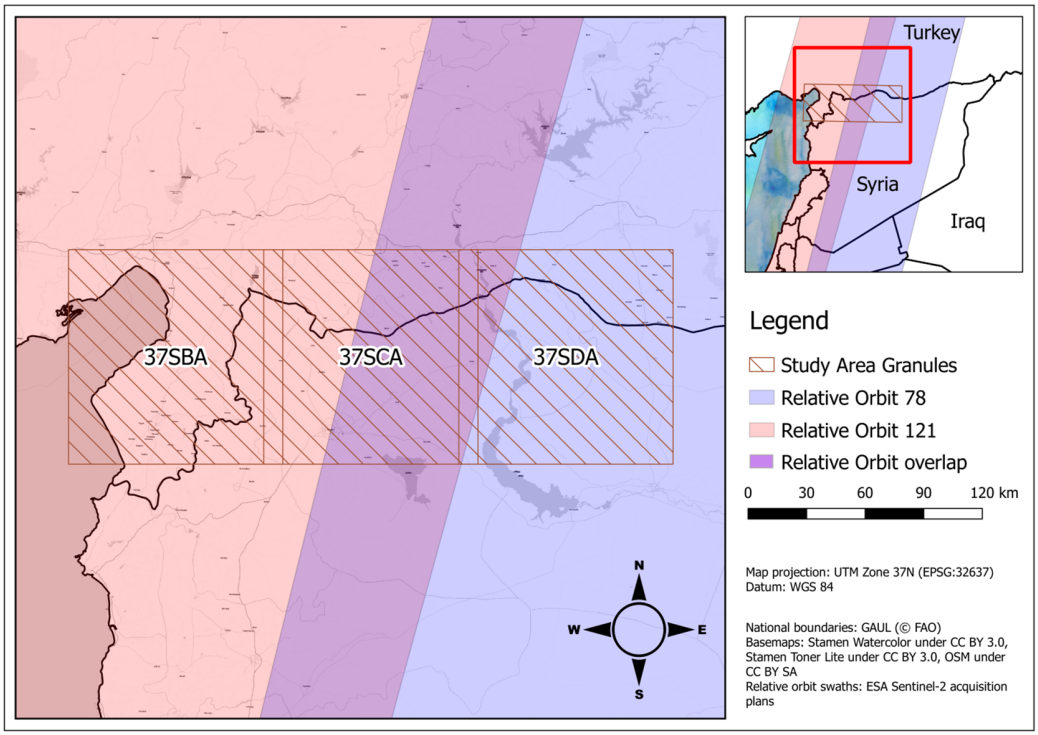

Fig. 2: Study area based on 3 overlapping granules (e.g. 37SBA, 37SCA, 37SDA provided in UTM Zone 37N) with overlapping Sentinel-2 relative acquisition orbits (modified from AUGUSTIN et al., 2018)

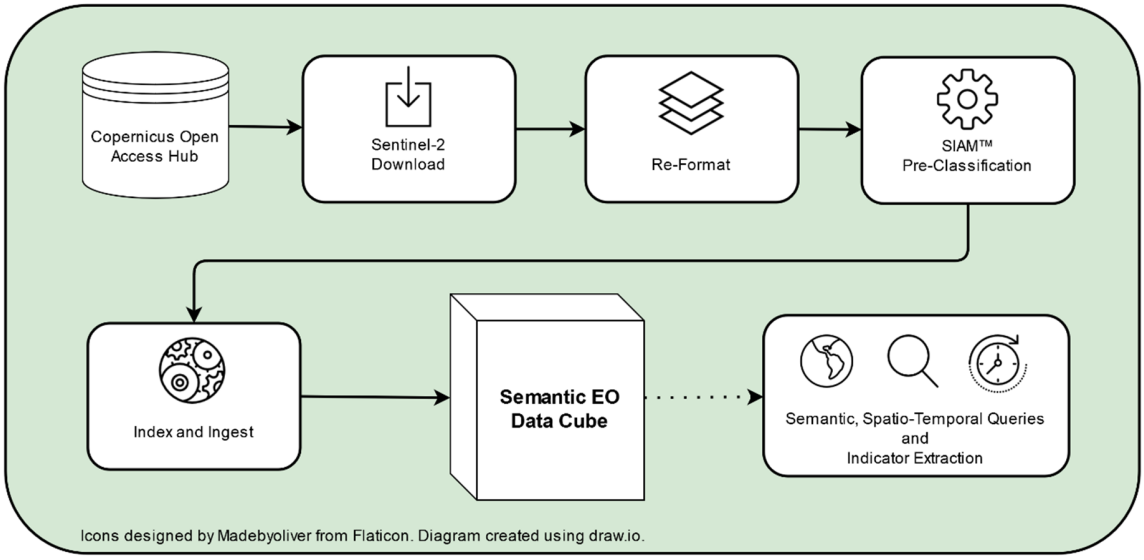

Fig. 3: Automated workflow overview from download to data cube incorporation (AUGUSTIN et al., 2018)

Spectral-based image pre-classification, as implemented by SIAM™, divides the feature space of a multi-spectral image into semantic semi-concepts using a knowledge-based approach, which is in contrast to data-driven approaches (e.g. supervised classification) (BARALDI et al. 2010; BARALDI 2011; BARALDI 2018). This could also be called descriptive colour naming, because the resulting semi-categories refer to similar pixels in terms of the multi-spectral information a pixel can offer. Assuming scenes are calibrated to a minimum of top-of-atmosphere reflectance, semi-concepts generated by SIAM™ are comparable and transferable between multiple images and optical sensors without any additional user-defined parameterisation (i.e. fully automatic). This is the only software used in the workflow that is closed-source.
All other software is open-source. The computing environment used for this implementation is a Red Hat Enterprise Linux 7 virtual machine with 16 virtual CPU at 2.5GHz clocking, 31GB RAM and 3TB of generic, all-use storage. The entire workflow is implemented using python in reproducible virtual environments using bash scripts run as automated cronjobs in Linux. The ODC technology is implemented with a PostgreSQL backend. Refer to AUGUSTIN et al. (2018) for more detailed information about the technical implementation.

2.2 Discussion and Example

The automated workflow stops at incorporation into the ODC implementation, resulting in a semantic EO data cube. This data cube can then be semi-automatically queried using existing Jupyter notebooks (i.e. existing blocks of interactive code), to construct various queries utilising the ODC's python API.
Concepts of reproducibility were strong drivers behind this implementation. The reproducibility of information extraction in this case is highly linked to the level of automation of information production, which is quite high. The methods and results of any query ought to be reproducible, given access to the same Sentinel-2 data, versions of the SIAM™ and ODC software, python computing environments, code and queries. Incorporating all Sentinel-2 data available for an

area including information layers generated through pre-classification in an implementation of the ODC can currently be considered as providing the data in an analysis-ready way. Due to the fully automated preparation of data from acquisition to ODC ingestion, this implementation can be considered highly repeatable. Given the automation developed in this work, a copy of this semantic data cube could be rebuilt in a similar computing environment in an estimated 5 days, assuming at least four parallel SIAM™ processes and stable downloads.

One example of a semantic query is for a normalised index of observed semi-concepts over time, e.g. for vegetation-like semi-concepts after having filtered out cloud-like semi-concepts (Fig. 4). As seen in Figure 5, such a query was conducted for nearly 3 years of data (28 June 2015 until 22 June 2018) over the entire spatial extent of the implementation (i.e. over 30,000km² with 10m pixels) (Fig. 5c and 5e). In addition, a query for the number of scenes available per pixel was conducted (Fig. 5a), as well as calculating the number of "clean pixels" by excluding cloud-like, snow-like, and unknown semi-concepts (i.e. relatively high reflectance in multiple spectral bands) (Fig. 5b). Each of these three massive queries took around 4 hours to complete in the given computing environment. Keep in mind, each query accessed the information layers belonging to 591 independent Sentinel-2 scenes. Figure 5d displays free and open OpenStreetMap (OSM) data to demonstrate the plausibility of the results, whereby multiple canals, buildings and even the border are visible in the absence of vegetation-like observations, and visible field structures coincide with the normalised vegetation occurrence output.

For such a massive query in terms of space, time and data volume, the number of scenes and clean pixels helps contextualise the heterogeneity of data underlying the results. What is lacking is a metric or analysis of distribution through time, which, for example, would make seasonal cloud-cover differences over large areas visible. Additionally, the simplified "clean pixel" calculation tends to exclude buildings (e.g. city of Aleppo), shallow water bodies (e.g. saline lake, Jabbūl, in the south), and other highly reflective, ambiguous surfaces. This could be overcome with more complex rules to take temporal (in)stability of semi-concepts into account, or their spatial neighbours (e.g. moving towards object detection and object-based methods in space and time).

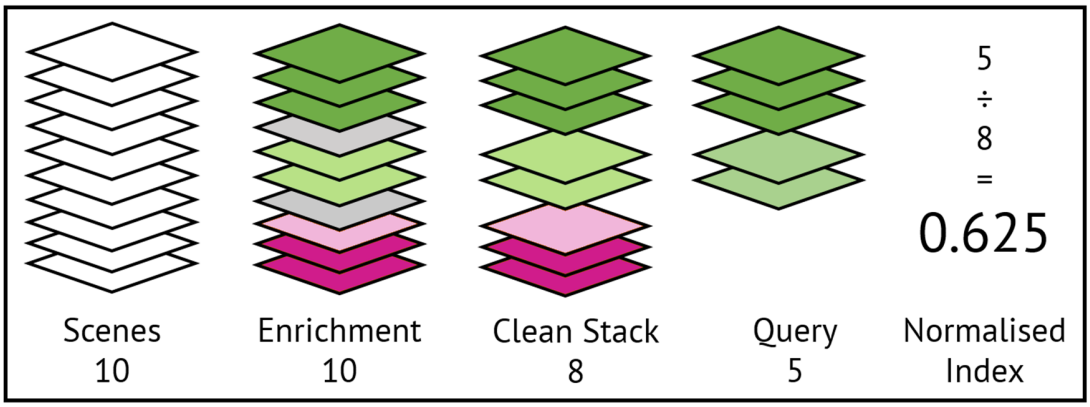

Fig. 4: This demonstrates how vegetation-like semi-concept occurrence aggregated over time is calculated. Green represents various vegetation-like semi-concepts, light-grey cloud-like, and pink/magenta anything else (e.g. bare soil). The clean stack contains observations excluding cloud-like semi-concepts following semantic-enrichment. The query selection contains only vegetation-like observations. Author's illustration

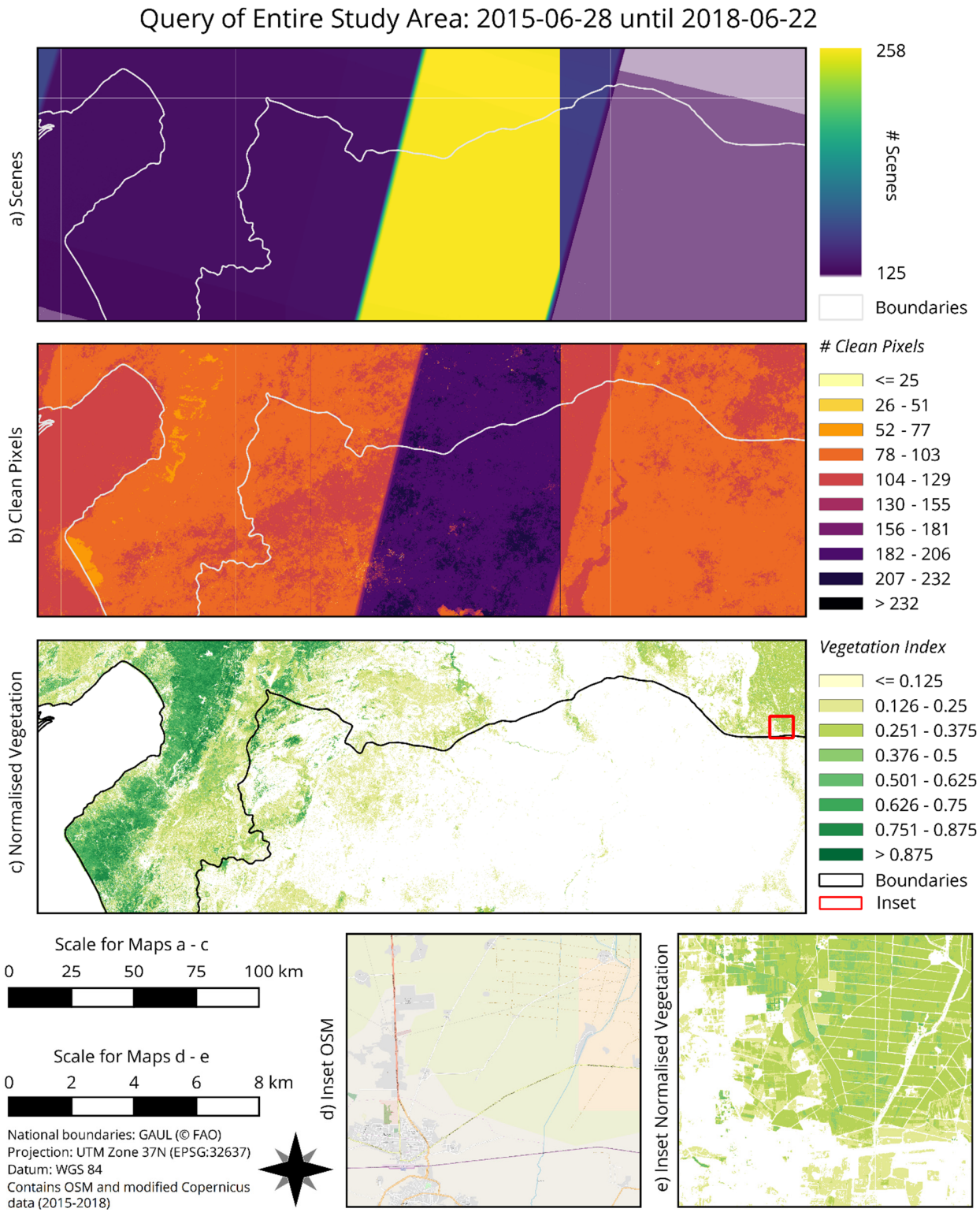

Fig. 5: Multiple semantic queries of the same spatio-temporal extent were conducted: (a) number of available scenes; (b) number of pixels excluding cloud-like, snow-like and unknown semi-concepts; (c) normalised vegetation index based on vegetation-like semi-concepts. A closer look of (c) identified with a red inset square is visible in (e). Existing OSM data for the same area and scale as (e) is visible in (d). Author's illustration

3 Looking Forward

The contribution and innovation demonstrated by this proof-of-concept implementation is the automated set-up of a semantic data cube. The semantic data cube contrasts to most other existing data cube implementations because it stores information and data together in an analysis-ready way, allowing ad-hoc multi-temporal, spatial and semantic queries, while supporting reproducible results. The generic, application-independent semantic enrichment utilised enables queries and EO-based indicator extraction for many thematic tasks. The semantic semi-concepts can be thought of as transferable, reproducible, sensor-agnostic building blocks for conducting further analysis. Given solid documentation on methods applied to generate output, reproducible results and repeatable analysis ought to be possible since the information layers (i.e. basis for semantic queries and analysis) continue to exist in the data cube and are stable concepts. This could be particularly relevant for supporting global initiatives (e.g. UN's sustainable development goals (UNITED NATIONS 2015b), Sendai Framework for Disaster and Risk Reduction (UNITED NATIONS 2015a)) because information is based on data collected independent of political borders and in a constant, relatively unbiased way.

However, the work presented here is not only a technical implementation towards developing indicators, but also an initial exploration of some of the challenges faced when working with dense time-series of EO data over larger spatial areas and timespans. These challenges directly result from qualities that are characteristic of big Earth data, i.e. their volume, velocity, and variety, but also varied methods of information generation and heterogeneity in underlying data and assumptions. This heterogeneity can encompass the distribution, variation, variability and uncertainty in space and time for all multi-temporal EO analysis and archives, especially when covering relatively large spatial extents.

This is a moment in time with unprecedented processing capabilities and free and open data availability. It is increasingly important to have meaningful, comprehensive and standardised methods to characterise and visualise uneven spatio-temporal distribution and coverage, uncertainty and variability as well as variation in data quality (e.g. cloud coverage) for different big Earth data sources and archives. Just because the data can possibly be considered unbiased in their regular, global acquisition does not mean that information generated from them are unbiased or will be used and understood in ways that are not misleading to different audiences. The more data that is incorporated in analysis and information generation, the more important it becomes to characterise the underlying data in terms of difference in quality and the previously mentioned characteristics.

Further research will include: (1) increasingly expressive or comprehensive rule-sets for queries taking spatial (e.g. neighbours, texture, objects), or temporal context (e.g. before and after, (dis)continuity, patterns like phenology) into account; (2) developing reliable indicators tested for agreement with existing sources that may be relevant to existing international initiatives; and (3) exploring methods and metrics to better assess the distribution, variation, variability and uncertainty inherent to dense, multi-temporal EO analysis and archives.

4 Acknowledgements

This work was supervised by Dr. Dirk Tiede and Martin Sudmanns, MSc, and financially supported by the Austrian Federal Ministry of Transport, Innovation and Technology (BMVIT) under the program "ICT of the Future" within the project SemEO (contract number: 855467). Continued work will be financially supported within the scope of the Sen2Cube.at project (contract number: 866016) funded by the Austrian Research Promotion Agency (FFG) under the Austrian Space Applications Programme (ASAP 14). None of this would have been possible without the contribution and support of Dr. Andrea Baraldi and access to his software, SIAM™. A free and openly available PDF copy of the complete master thesis this contribution is based on is available at https://github.com/augustinh22/msc_markdown/ in the output directory.

5 References

AUGUSTIN, H., SUDMANNS, M., TIEDE, D. & BARALDI, A., 2018: A Semantic Earth Observation Data Cube for Monitoring Environmental Changes during the Syrian Conflict. GI_Forum 2018, **1**, 214-227.

BARALDI, A., 2018: Satellite Image Automatic Mapper™ System and Products Description. Retrieved 29 January 2018, from http://siam.andreabaraldi.com/content/Documentation/SIAM_Report_BACRES_v1.18.pdf

BARALDI, A., DURIEUX, L., SIMONETTI, D., CONCHEDDA, G., HOLECZ, F. & BLONDA, P., 2010: 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 Transactions on Geoscience and Remote Sensing, **48**(3), 1299-1325.

BARALDI, A., 2011: Satellite Image Automatic Mapper™ (SIAM™) - A Turnkey Software Executable for Automatic Near Real-Time Multi-Sensor Multi-Resolution Spectral Rule-Based Preliminary Classification of Spaceborne Multi-Spectral Images. Recent Patents on Space Technology, **1**(2), 81-106.

DRUSCH, M., DEL BELLO, U., CARLIER, S., COLIN, O., FERNANDEZ, V., GASCON, F., … BARGELLINI, P., 2012: Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services. Remote Sensing of Environment, **120**, 25-36.

EUROPEAN SPACE AGENCY, 2017: Sentinel High Level Operations Plan (HLOP): COPE-S1OP-EOPG-PL-15-0020. Retrieved from https://earth.esa.int/documents/247904/685154/Sentinel_High_Level_Operations_Plan

SOILLE, P., BURGER, A., DE MARCHI, D., KEMPENEERS, P., RODRIGUEZ, D., SYRRIS, V. & VASILEV, V., 2018: A versatile data-intensive computing platform for information retrieval from big geospatial data. Future Generation Computer Systems, **81**, 30-40.

TIEDE, D., LÜTHJE, F. & BARALDI, A., 2014: Automatic post-classification land cover change detection in Landsat images: Analysis of changes in agricultural areas during the Syrian crisis. Publikationen der Deut. Gesellschaft für Photogrammetrie, Fernerkundung u. Geoinformation e.V., Band **23**, Beitrag 191.

UNITED NATIONS, 2015a: Resolution 69/283: Sendai Framework for Disaster Risk Reduction 2015–2030. United Nations General Assembly. Retrieved from https://undocs.org/A/RES/69/283

UNITED NATIONS, 2015b: Resolution 70/1: Transforming our world: the 2030 agenda for sustainable development. United Nations General Assembly. Retrieved from https://undocs.org/A/RES/70/1

WULDER, M. A., MASEK, J. G., COHEN, W. B., LOVELAND, T. R. & WOODCOCK, C. E., 2012: Opening the archive: How free data has enabled the science and monitoring promise of Landsat. Remote Sensing of Environment, **122**, 2-10.