From Copernicus Big Data to Extreme Earth Analytics

Visionary Paper

Manolis Koubarakis[1], Konstantina Bereta[1], Dimitris Bilidas[1], Konstantinos Giannousis[1], Theofilos Ioannidis[1], Despina-Athanasia Pantazi[1], George Stamoulis[1], Seif Haridi[2], Vladimir Vlassov[2], Lorenzo Bruzzone[3], Claudia Paris[3], Torbjørn Eltoft[4], Thomas Krämer[4], Angelos Charalabidis[5], Vangelis Karkaletsis[5], Stasinos Konstantopoulos[5], Jim Dowling[6], Theofilos Kakantousis[6] Mihai Datcu[7], Corneliu Octavian Dumitru[7], Florian Appel[8], Heike Bach[8], Silke Migdall[8], Nick Hughes[9], David Arthurs[10], Andrew Fleming[11]

[1] National and Kapodistrian University of Athens, [2] KTH Royal Institute of Technology, Stockholm
[3] University of Trento, [4] University of Tromsø, [5] National Center for Scientific Research - Demokritos
[6] LogicalClocks, [7] German Aerospace Center, [8] VISTA Remote Sensing in Geosciences GmbH
[9] Norwegian Meteorological Institute, [10] Polar View, [11] British Antarctic Survey
koubarak@di.uoa.gr

ABSTRACT

Copernicus is the European programme for monitoring the Earth. It consists of a set of systems that collect data from satellites and in-situ sensors, process this data and provide users with reliable and up-to-date information on a range of environmental and security issues. The data and information processed and disseminated puts Copernicus at the forefront of the big data paradigm, giving rise to all relevant challenges, the so-called 5 Vs: volume, velocity, variety, veracity and value. In this short paper, we discuss the challenges of extracting information and knowledge from huge archives of Copernicus data. We propose to achieve this by scale-out distributed deep learning techniques that run on very big clusters offering virtual machines and GPUs. We also discuss the challenges of achieving scalability in the management of the extreme volumes of information and knowledge extracted from Copernicus data. The envisioned scientific and technical work will be carried out in the context of the H2020 project ExtremeEarth which starts in January 2019.

1 INTRODUCTION

Copernicus is the European programme for monitoring the Earth. It consists of a set of systems that collect data from satellites and in-situ sensors, process this data and provide users with reliable and up-to-date information on a range of environmental and security issues. The Earth observation satellites that provide the data of Copernicus are the *Sentinels*, which are developed for the specific needs of the Copernicus programme, and the *contributing missions*, which are operated by national, European or international organizations. The access to Sentinel data is regulated by EU law and it is full, open and free. Information extracted from Copernicus data is made available to users through the *Copernicus services* addressing six thematic areas: land, marine, atmosphere, climate, emergency and security.

The data and information processed and disseminated puts Copernicus at the forefront of the big data paradigm, giving rise to all relevant challenges, the so-called *5 Vs*, discussed below.

Volume: The repository of Sentinel products managed by the European Space Agency (ESA) has so far published more than 5 million products, and it has more than 100 thousand users who have downloaded more than 50 PB of data since the start of the operations of the system. This volume will increase in the following years, as new Sentinel satellites are launched.

Velocity: Copernicus data has to be delivered and processed in a short time frame to allow the provision of 24/7 information to users requiring fast responses. By the end of 2016, 6 TB of data were generated and 100 TB of data were disseminated every day from the Sentinel product repository. These rates will increase in forthcoming years as new Sentinel satellites are launched.

Variety: The Sentinel satellites have different types of sensors (e.g., radar and optical) and different levels of processing (from raw data to advanced products). Moreover, datasets used for geospatial applications can be not only satellite data but also aerial imagery, in-situ data and other collateral information (e.g., public government data). This wealth of data is processed by Earth Observation actors to extract information and knowledge. This *information and knowledge is also big* and similar big data challenges apply. For example, 1PB of Sentinel data may consist of about 750.000 datasets which, when processed, about 450TB of content information and knowledge (e.g., classes of objects detected) can be generated.

Veracity: Decision-making and operations require reliable sources. Thus, assessing the quality of the data is important for the whole information extraction chain.

Value: The extraction of information from the Copernicus data has direct economic benefits for Europe. Several economic studies have concluded that the Copernicus programme has the potential to significantly impact job creation, innovation and growth. The Copernicus Market report of 2016 estimates that the overall investment in Copernicus will reach EUR 7.4 billion in the years 2008-2020, while the cumulative economic value generated by it in the same period will be around EUR 13.5 billion, and it will support 28.030 job years in the Earth Observation sector.

An important activity related to Copernicus is the *thematic exploitation platforms (TEPs)* of the European Space Agency (ESA). A TEP is a collaborative, virtual work environment addressing a class of users and providing access to EO data, algorithms and computing/networking resources required to work with them, through one coherent interface. The fundamental principle of the TEPs is to move the user to the data and tools as opposed to the traditional approach of downloading, replicating, and exploiting data "at home". Now the user community is present and visible in the platform, involved in its governance and and enabled to share and collaborate. There are currently 7 TEPs addressing the following application areas: coastal, forestry, hydrology, geohazards, polar, urban themes, and food security.

Another important development in the context of Copernicus is the implementation of five *Copernicus Data and Information Access Services (DIAS).* The European Commission has awarded in December 2017 four contracts to industrial consortia for the development of four cloud-based platforms for Copernicus DIAS. The fifth DIAS is built by EUMETSAT in collaboration with Mercator Ocean and the European Centre for Medium-Range Weather Forecasts. Like the TEPs, these five platforms will also bring computing resources close to the data and enable an even greater commercial exploitation of Copernicus data.

Although the TEPs and DIAS activities funded by ESA have been welcomed by the EO data user community, they both have a *significant disadvantage:* they target users that are experts in EO data and technologies, and ignore the myriad of software developers that might not be experts in EO but still have a lot to gain by integrating EO data in their applications. Therefore, *opening up the TEPs and DIASs* by extracting information and knowledge hidden in the data, publishing this information and knowledge using *linked data technologies*, and interlinking it with data in other TEPs and DIASs and other non-EO data, information and knowledge can be an important way of making the development of downstream applications easy for both EO and non-EO experts.

In the last few years, there have been four highly successful European research projects that have pursued this idea: the FP7 projects TELEIOS , LEO and Melodies , and the on-going project Copernicus App Lab . These projects pioneered the use of linked geospatial data in the EO domain, and demonstrated the potential of linked data and semantic web technologies by developing prototype environmental and business applications. However, *none of these projects has faced the challenges of big data, information and knowledge that users and application developers are facing in the context of the Copernicus program.* The above four projects have developed tools for knowledge discovery and data mining from satellite images and related geospatial data sets, as well as tools for linked geospatial data integration, querying and analytics. However, *none of these tools scales to the many PBs of data, information and knowledge present in the Copernicus context.* For example, the state-of-the art geospatial and temporal RDF store Strabon implemented in TELEIOS [15] can only handle up to 100 GBs of point data and still be able to answer simple geospatial queries (selections over a rectangular area) efficiently (in a few seconds). Competitor systems like GraphDB by company OntoText performs similarly [2]. If the complexity of geometries in the dataset increases (i.e., we have multi-polygons), not even the aforementioned performance can be achieved for both Strabon and GraphDB.

In addition, contrary to multimedia images, for which highly scalable Artificial Intelligence techniques based on deep neural network architectures have been developed by big North American companies such as Google and Facebook recently [7, 8], *similar architectures for satellite images, that can manage the extreme scale and characteristics of Copernicus data, do not exist today.* The deep neural network architectures can classify effectively and efficiently multimedia images because they have been trained using extremely large benchmark datasets consisting of millions of images (e.g., ImageNet) and have utilized the power of big data, cloud and GPU technologies. *Training datasets consisting of millions of data samples in the Copernicus context do not exist today* and published deep learning architectures for Copernicus satellite images typically run using one GPU and do not take advantage of recent advances like distributed scale-out deep learning [8].

2 MAIN OBJECTIVE AND TECHNICAL CHALLENGES

The main objective of ExtremeEarth is to go beyond the four projects mentioned above by developing *extreme Earth analytics techniques and technologies that scale to the PBs of big Copernicus data, information and knowledge, and applying these technologies in two of the ESA TEPs: Food Security and Polar.* The technologies to be developed will extend the HOPS data platform [9, 12, 13, 17] to offer unprecedented scalability to extreme data volumes and scale-out distributed deep learning for Copernicus data. The extended HOPS data platform will run on a DIAS selected after the project starts and will be available as open source to enable its adoption by the strong European Earth Observation downstream services industry.

The detailed scientific and technical challenges of ExtremeEarth are the following.

Challenge C1. To develop scalable deep learning and extreme earth analytics techniques for Copernicus big data. The constellations of Sentinel-1/2/3 satellites have the important capability to acquire long time series of multispectral and Synthetic Aperture Radar (SAR) images where the temporal dimension plays a very important role for the characterization of the information content of the image (e.g., land cover or sea ice) and its dynamics. Moreover, this results in the availability of very large archives of images covering long time periods. Another key aspect of the Sentinel missions is the multimodal structure of the platforms. Different kinds of sensors (radar, optical, multi/multispectral) are available and can be used in synergy. Each modality provides specific information that can be used to cope with the limitations of another. ExtremeEarth will advance the state of the art in this area [20] by developing distributed scale-out deep learning techniques for the classification of remote sensing images based on architectures that can effectively exploit the spatial, spectral, temporal and multimodal properties of Sentinel data.

Two deep learning architectures for the classification of Sentinel remote sensing images will be developed; one for determining crop boundaries and type, and one for sea ice mapping. These will be used in the two applications described in the Challenges A1 and A2 below. The developed algorithms will be supported by a scale-out open-source platform for distributed deep learning and big data, based on the HOPS data platform [12].

Challenge C2. To develop very large training datasets for deep learning architectures targeting the classification of Sentinel images. In deep learning architectures, the availability of large amounts of high quality training data is equally important to the learning models. Computer vision and other image processing areas have developed during the last few years huge training data sets (e.g., ImageNet) consisting of many millions of objects. Satellite remote sensing is largely lacking this development since the complexity and physical meaning of the sensor data make the generation of training datasets much more complex. Moreover, from an operational viewpoint it is not feasible to assume the availability of enough ground truth or annotated labeled data for training a deep network. In this area, the largest benchmark dataset is Eurosat which was proposed recently [11]. It uses Sentinel 2 data and covers 13 different spectral bands and 10 land cover classes with a total of 27,000 labeled images.

In ExtremeEarth, we will develop tools to generate EO training datasets by enlarging existing datasets currently in development by the German Aerospace Center [4, 18] and by leveraging existing cartographic/thematic products which are now available at continental or planetary scale (e.g., OpenStreetMap). Two training datasets consisting of millions of samples will be developed aimed at the two deep learning architectures of Challenge C1. The datasets will be published as open source to be used by the whole Remote Sensing community.

Challenge C3. To develop techniques and tools for linked geospatial data querying, federation and analytics that scale to big Copernicus data, information and knowledge. The paradigm of linked geospatial data and relevant technologies has been pioneered by previous projects TELEIOS, LEO, Melodies and Copernicus App Lab mentioned above. The technologies developed include state-of-the-art systems for transforming geospatial data into RDF (GeoTriples [16]), interlinking with other geospatial data sources (geospatial/temporal extensions of Silk [21]), visualizing (Sextant [5]), querying, federating (Semagrow [3]) and performing data analytics (Strabon [15] and Ontop-spatial [1]). These systems are open-source and they currently represent the international state-of-the-art in the area of linked geospatial and EO data [6, 14]. In ExtremeEarth, the systems GeoTriples and Strabon will be re-engineered so that they scale to big linked geospatial data and extreme geospatial analytics. In addition, the JedAI linking framework [19] will be extended to enable the scalable discovery of geospatial relations in big geospatial RDF data sources. Finally, the engine Semagrow will be extended so that it can manage efficiently federations of big geospatial data sources and answer extreme geospatial analytical queries. To achieve the required extreme scalability, we will develop these three systems on top of the highly scalable HOPS data platform [12].

Challenge C4. To extend the capabilities for EO data discovery and access with semantic catalogue services that scale to the big data, information and knowledge of Copernicus. Currently, Copernicus data catalogues (e.g., the Copernicus Open Access Hub or the catalogues of the various TEPs) allow a user to access data by drawing an area of interest on the map and specifying search parameters such as sensing date, mission, satellite platform, product type etc. The new semantics-based catalogue we will develop in ExtremeEarth will expose the knowledge hidden in Sentinel satellite images and related data sets, and will allow a user to ask sophisticated queries such as “How many icebergs were embedded in the Norske ÃŸer Ice Barrier at its maximum extent in 2017?” which currently cannot be answered by the catalogue of the Norwegian Meteorological Institute, although all this knowledge is available in the Sentinel archive and related European data sets. We will demonstrate how to develop semantics-based catalogues and how to implement them efficiently for the extreme scale of the Copernicus context using the advances of ExtremeEarth in the area of big linked geospatial data discussed in Challenge C3 above. Two semantic catalogues (one for each TEP) will be developed operating in the selected DIAS platform and scaling to trillions of metadata records.

Challenge C5. To integrate the big data and extreme earth analytics technologies of Challenges C1-C4 in the HOPS data platform and deploy them in the selected DIAS and the two TEPs. The ExtremeEarth technologies presented above will be implemented and evaluated in the elastic cloud environment of the startup LogicalClocks participating in the consortium. This cloud environment is managed by LogicalClocks and currently provides the HOPS data platform [9, 12, 13, 17] together with significant storage, compute and GPU resources that will be made available to the project. Copernicus data will also be made available in the same environment and will be used to develop the ExtremeEarth technologies. The HOPS data platform provides services to move the processing to where the data is and it is based on a cloud computing platform-as-a-service approach. HOPS supports state-of-the-art parallel processing on big data with Apache Spark and deep learning with TensorFlow/Keras , as well as distributed deep learning using TensorFlow’s distribution strategies, including collective allreduce and parameter server . HOPS also provides its own libraries for parallel deep learning experiments (hyperparameter search and model-architecture search). Once the ExtremeEarth technologies are integrated in HOPS, they will be deployed in the two TEPs and the selected DIAS.

The technologies discussed above will be demonstrated in two application areas: Food Security and Polar addressed by the relevant ESA TEPs. The challenges to be addressed in these two applications are the following.

Challenge A1. To develop high resolution water availability maps for agricultural areas allowing a new level of detail for wide-scale irrigation support. The maps will be available as linked data together with other geospatial layers (e.g., OpenStreetMap, field boundaries, crop types etc.) and made available to farmers. ExtremeEarth will tackle the combination of hydrological and agricultural monitoring, both in the sense that the thematic areas will be brought together, but also in the sense that the federation of the data sources and the cloud platforms used for the processing will be integrated. Both TEPs are already up and running, with the Food Security TEP, as the youngest of the TEPs, still in its second development phase, and the Polar TEP being in pre-operational mode. On both platforms, the first pre-processing chains for the information needed in ExtremeEarth are already running. These will be the baseline for bringing the two applications together in one combined application for irrigation. In practice this means that processing has to be widened to include whole watersheds (or catchment areas), to include all necessary Copernicus satellite input data from radar and optical imagery and to span the whole year instead of just the winter season or vegetation period. Additionally, scalable deep learning techniques discussed in Challenge C1 will be used to derive field boundaries and crop types, making it possible for the processing chains to include this information as linked data on a large scale (formerly, this information was only available at farm level). This will allow crop type specific deduction of crop variables, and thus a higher degree of accuracy for each field. The information generated with the ExtremeEarth approach will then be fed into the PROMET model [10] to provide high resolution (10m) water availability maps for the agricultural area in the whole watershed, allowing a new level of detail for wide-scale irrigation support. This application, which combines different TEPs, different Earth observation types and remotely sensed information and land surface modelling can be seen as a blueprint for further such applications, where the focus will lie on a federation of specialized knowledge and working environments (data and workflows). This type of federation of TEPs with methods, tools and data specialised for their topic rather than one broad platform for everything is seen by us as the way into the future.

Challenge A2. To produce high resolution ice maps from massive volumes of heterogeneous Copernicus data. The maps will be made *available as linked data and will be combined with other information such as sea surface temperature and wind information for informing maritime users.* The anticipated economic development of the Arctic, partially driven by reductions in sea ice cover, will see an increase in maritime shipping activity. High quality, timely and reliable information about sea ice and iceberg conditions is vital to ensure that vessels navigate efficiently and safely with minimal risk to the environment. This information is required by vessels in many sectors, including cargo transport, fisheries, tourism, research vessels, resource exploration and extraction, destination shipping and national coast guard vessels.

The functionality provided by both the DIAS infrastructure and the ESA Polar TEP as part of the ExtremeEarth infrastructure are well suited to answering the challenges of this application. Access to the required data interfaces is already established for some data sources in Polar TEP and further work to reinforce this will happen with integration of the DIAS infrastructure. Effort will be required to ensure data access is optimised for these purposes, for example ensuring near-real-time access to all required datasets. Building on the Polar TEP and DIAS as part of the ExtremeEarth infrastructure will also provide access to compute resources for processing. Since this is potentially going to be a significant processing load, but for limited periods of time as data is acquired and becomes available, then processing resources will need to be on demand and scalable to ensure efficiency. This will be achived by building on top of the HOPS data platform. Integration of established delivery systems into the ExtremeEarth infrastructure will support delivery of information products to polar users, such as tourist ships and fishing vessels operating in ice infested waters. This will include systems for information delivery and visualisation such as the Polar Code Decision Support System (PCDSS) which is currently being developed by company Polar View. PCDSS is designed to be used over restricted communication links, to bridge between the service production and users onboard ships in the Polar Regions. The scalable deep learning algorithms for sea ice classification, discussed in Challenge C1 above, will be integrated in the HOPS data platform to produce high resolution ice maps from massive volumes of heterogeneous Copernicus data. The aim is to deliver sea ice concentration and type maps, displaying stage of development (in accordance with the World Meteorological Organization - WMO Sea Ice Nomenclature), including fraction of leads and ridges, over the Polar Regions, at a resolution of 1 km or better.

3 THE EXTREMEEARTH CONSORTIUM

ExtremeEarth brings together a consortium of two companies leading the activities of the Food Security and Polar TEPs (VISTA and Polar View), one organization specializing in polar science for the Arctic and the Antarctic (British Antarctic Survey), one company specializing in big data, analytics and deep learning technologies (LogicalClocks), the German Aerospace Center with its TerraSAR-X satellite and expertise in SAR and multispectral EO (DLR), five top European academic institutions specializing in big data, linked data, Artificial Intelligence, deep learning and extreme earth analytics (National and Kapodistrian University of Athens, University of Trento, University of Tromsø, KTH and DLR), one research institute specializing in big data and Artificial Intelligence (National Center for Scientific Research - Demokritos), and one research institute specializing in big data, high performance computing applications, and provision of Metocean information, including sea ice (Norwegian Meteological Institute). The consortium is led by the National and Kapodistrian University of Athens.

ExtremeEarth starts in January 1, 2019 and will have a duration of 3 years. The project consortium is fully aware of the huge challenges that lay in front of us, and it is looking forward to meet them!

4 CONCLUSIONS

We have presented the vision of Horizon 2020 European project ExtremeEarth addressing the challenges of how to extract information and knowledge from the PBs of satellite data of the Copernicus programme using deep learning techniques, how to manage this information and knowledge efficiently using the HOPS data platform, how to develop two applications with economic and environmental importance (Food Security and Polar), and how to deploy these applications on the two relevant ESA TEPs and DIAS.

ACKNOWLEDGMENTS

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825258.

REFERENCES

[1] K. Bereta and M. Koubarakis. 2016. Ontop of Geospatial Databases. In *ISWC*.

[2] K. Bereta, M. Koubarakis, S. Manegold, G. Stamoulis, and B. Demir. 2018. From Big Data to Big Information and Big Knowledge: The Case of Earth Observation Data. In *CIKM*.

[3] A. Charalambidis, A. Troumpoukis, and S. Konstantopoulos. 2015. SemaGrow: optimizing federated SPARQL queries. In *SEMANTICS*.

[4] C. Dumitru, G. Schwarz, and M. Datcu. 2018. SAR Image Land Cover Datasets for Classification Benchmarking of Temporal Changes. *J-STARS* 11 (2018).

[5] C. Nikolaou et al. 2015. Sextant: Visualizing time-evolving linked geospatial data. *J. Web Sem.* 35 (2015), 35–52.

[6] M. Koubarakis et al. 2016. Managing Big, Linked, and Open Earth-Observation Data: Using the TELEIOS/LEO software stack. *IEEE Geoscience and Remote Sensing Magazine* 4, 3 (2016).

[7] O. Russakovsky et al. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision* 115, 3 (2015), 211–252.

[8] P. Goyal et al. 2017. Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour. *CoRR* (2017).

[9] S. Niazi et al. 2017. HopsFS: Scaling Hierarchical File System Metadata Using NewSQL Databases. In *FAST*.

[10] T. Hank, H. Bach, and W. Mauser. 2015. Using a Remote Sensing-Supported Hydro-Agroecological Model for Field-Scale Simulation of Heterogeneous Crop Growth and Yield: Application for Wheat in Central Europe. *Remote Sensing* 7, 4 (2015), 3934–3965.

[11] P. Helber, B. Bischke, A. Dengel, and D. Borth. 2018. Introducing Eurosat: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification. In *IGARSS*.

[12] M. Ismail, E. Gebremeskel, T. Kakantousis, G. Berthou, and J. Dowling. 2017. Hopsworks: Improving User Experience and Development on Hadoop with Scalable, Strongly Consistent Metadata. In *ICDCS*.

[13] M. Ismail, S. Niazi, M. Ronström, S. Haridi, and J. Dowling. 2017. Scaling HDFS to more than 1 million operations per second with HopsFS. In *CCGRID*.

[14] M. Koubarakis, K. Bereta, G. Papadakis, D. Savva, and G. Stamoulis. 2017. Big, Linked Geospatial Data and Its Applications in Earth Observation. *IEEE Internet Computing* 21, 4 (2017).

[15] K. Kyzirakos, M. Karpathiotakis, and M. Koubarakis. 2012. Strabon: A Semantic Geospatial DBMS. In *ISWC*.

[16] K. Kyzirakos, D. Savva, I. Vlachopoulos, A. Vasileiou, N. Karalis, M. Koubarakis, and S. Manegold. 2018. GeoTriples: Transforming geospatial data into RDF graphs using R2RML and RML mappings. *J. Web Sem.* (2018).

[17] S. Niazi, M. Ronström, S. Haridi, and J. Dowling. 2018. Size Matters: Improving the Performance of Small Files in Hadoop. In *Middleware*.

[18] A. Oca, R. Bahmanyar, N. Nistor, and M. Datcu. 2017. Earth observation image semantic bias: A collaborative user annotation approach. *J-STARS* 10 (2017).

[19] G. Papadakis, K. Bereta, T. Palpanas, and M. Koubarakis. 2017. Multi-core Meta-blocking for Big Linked Data. In *SEMANTICS*.

[20] C. Persello and L. Bruzzone. 2014. Active and Semisupervised Learning for the Classification of Remote Sensing Images. *IEEE Trans. Geoscience and Remote Sensing* 52, 11 (2014), 6937–6956.

[21] P. Smeros and M. Koubarakis. 2016. Discovering Spatial and Temporal Links among RDF Data. In *LDOW*.