From Big Data to Big Information and Big Knowledge: the Case of Earth Observation Data*

Konstantina Bereta
Dept. of Informatics and Telecommunications, National and Kapodistrian University of Athens, Greece
konstantina.bereta@di.uoa.gr

Manolis Koubarakis
Dept. of Informatics and Telecommunications, National and Kapodistrian University of Athens, Greece
koubarak@di.uoa.gr

Stefan Manegold
Database Architectures Group, CWI, The Netherlands
Stefan.Manegold@cwi.nl

George Stamoulis
Dept. of Informatics and Telecommunications, National and Kapodistrian University of Athens, Greece
gstam@di.uoa.gr

Begüm Demir
Faculty of Electrical Engineering and Computer Science, Technische Universität Berlin, Germany
demir@tu-berlin.de

CCS CONCEPTS

• **Information systems → Database management system engines**; **Query languages for non-relational engines**; **Specialized information retrieval**; *Web searching and information discovery*;

KEYWORDS

Semantic web, linked geospatial data, Earth observation data, Copernicus program

ACM Reference Format:
Konstantina Bereta, Manolis Koubarakis, Stefan Manegold, George Stamoulis, and Begüm Demir. 2018. From Big Data to Big Information and Big Knowledge: the Case of Earth Observation Data. In *The 27th ACM International Conference on Information and Knowledge Management (CIKM '18), October 22–26, 2018, Torino, Italy.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3269206.3274270

1 INTRODUCTION

Some particularly important rich sources of open and free big geospatial data are the *Earth observation (EO)* programs of various countries such as the Landsat program of the US and the Copernicus programme of the European Union. EO data is a paradigmatic case of *big data* and the same is true for the *big information* and *big knowledge* extracted from it. EO data (satellite images and in-situ data), and the information and knowledge extracted from it, can be utilized in many applications with financial and environmental impact in areas such as emergency management, climate change, agriculture and security. This potential has not been fully realized up to now, because EO data and the information and knowledge extracted from it "is hidden" in various archives operated by NASA, ESA and national space agencies. Therefore, a user that would like to develop an application needs to search in these archives, discover the needed data and information and integrate it in his application. In this tutorial we show how to "break these silos open" by publishing their data as RDF, enable their discovery by modern search engines, interlink it with other relevant data, and make it freely available on the Web to enable the easy development of geospatial applications.

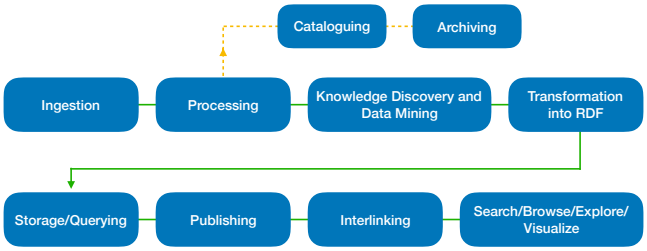

Figure 1: A data science pipeline for linked EO data

The tutorial will start by explaining why EO data is a paradigmatic case of big data giving rise to all relevant challenges, the so-called 5 Vs: *volume, velocity, variety, veracity* and *value*. Examples of big EO data, information and knowledge will be given for the case of the Copernicus programme of the European Union (http://copernicus.eu/).

The life of big EO data starts with its generation in the ground segment of a satellite mission. The management of this so-called *payload data* is an important activity of the ground segments of satellite missions. Figure 1 gives a high-level view of the data science pipeline for linked EO data that we have developed and used in the projects TELEIOS (http://www.earthobservatory.eu/), LEO (http://www.linkedeodata.eu/), MELODIES (http://www.melodiesproject.eu/), Copernicus App Lab (http://www.app-lab.eu/) and Big Data Europe (https://www.big-data-europe.eu/).

*This work has been supported by Horizon 2020 projects Copernicus App Lab (730124) and BigEarth (759764).

CIKM '18, October 22–26, 2018, Torino, Italy

ACM ISBN 978-1-4503-6014-2/18/10.
https://doi.org/10.1145/3269206.3274270

The pipeline starts with EO datasets in various formats that are made freely available in the archives of space agencies like ESA and NASA, and ends with the deployment of an interactive visual application that uses EO data, information and knowledge together with other collateral data (e.g., open government data, closed enterprise data, model data etc.) using linked data technologies. Each stage of the pipeline and its associated techniques and software tools developed by the presenters will be surveyed in the tutorial. The tutorial will also pay articular attention to the latest developments of the BigEarth project (http://bigearth.eu/) that has been funded by the European Research Council (ERC) and aims at providing a powerful capability to quickly and accurately access and extract vital information for observing the Earth from big EO archives. Related work by other researchers will also be covered in depth. Finally, open problems and directions for future research in this area will also be discussed.

Some of our work in the last 8 years in this area, which is the basis for a large part of this tutorial, is summarized in [1, 2]. The relevant tools supporting the linked EO data pipeline (Strabon, Ontop-spatial, GeoTriples, Silk and Sextant) are available as open source on the Web site http://kr.di.uoa.gr/.

2 INTENDED AUDIENCE

The tutorial is aimed at database, information retrieval and knowledge management researchers who would like to understand the state of the art and open problems in data science pipelines for EO data and linked geospatial data, and practitioners who would like to develop applications using existing tools. The tutorial assumes familiarity with RDF, SPARQL and geospatial data.

3 PRESENTERS

Konstantina Bereta is a Research Associate in the Dept. of Informatics and Telecommunications, National and Kapodistrian University of Athens, and she holds a BSc and MSc from the same department. She is also a PhD candidate under the supervision of Prof. Manolis Koubarakis (expected date of graduation: Fall 2018). She has worked as a scientific programmer and research associate in several EU FP7 projects. Her research interests focus in the areas of spatiotemporal databases, Semantic Web and Cloud Computing.

Manolis Koubarakis is a Professor in the Dept. of Informatics and Telecommunications, National and Kapodistrian University of Athens. He previously held positions at the Dept. of Electronic and Computer Engineering, Technical University of Crete (Assistant and Associate Professor), the Dept. of Informatics, University of Athens (Visiting Researcher), the Dept. of Computation, UMIST (now University of Manchester) (Lecturer) and the Dept. of Computing, Imperial College, London (Research Associate). He has published more than 180 papers that have been widely cited in the areas of Artificial Intelligence (especially Knowledge Representation), Databases, Semantic Web and Linked Data. In 2015, he was elected Fellow of the European Association of Artificial Intelligence (EurAI). He has served in the program committee of various international conferences and workshops, and he has organized various international events. He has attracted more than 6M Euros in funding from the European Commission, the Greek General Secretariat from Research and Technology, the European Space Agency and industry sources.

Stefan Manegold is the lead of the Database Architectures group of CWI and a Professor in Leiden University. He is a nationally and internationally recognized expert in system-oriented database research. He is particularly known for his pioneering work on hardware-conscious database technology, and for disseminating his research via the open-source columnar analytical database management system MonetDB, which is widely used in academia and business. Dr. Manegold's research is focused on bridging the gap between database architectures and demanding applications areas, such as large-scale data analytics (Big Data), data intensive scientific discovery (eScience), and semantic web. His expertise comprises database architectures, query processing algorithms, and data management technology, with a particular focus on hardware-conscious algorithms and data structures, query optimization, scalability, performance, benchmarking and testing.

George Stamoulis is a Research Associate in the Dept. of Informatics and Telecommunications, National and Kapodistrian University of Athens, and a PhD candidate under the supervision of Prof. Koubarakis. He holds a BSc and MSc from the Department of Informatics and Telecommunications of the National and Kapodistrian University of Athens. His research interests focus in the areas of Semantic Web, Data Visualization and Integration and User Interfaces.

Begüm Demir is a Professor and Chair of the Remote Sensing Image Analysis (RSiM) group at the Faculty of Electrical Engineering and Computer Science, Technische Universität Berlin (TU Berlin), Germany. Before joining to TU Berlin, she was an Assistant Professor at the Department of Computer Science and Information Engineering, University of Trento, Italy, from 2013 to 2017 while in 2017 she became an Associate Professor at the same department. Her main research interests include machine learning and big data management with applications to remote sensing image analysis. She was a recipient of an ERC Starting Grant with the project "BigEarth-Accurate and Scalable Processing of Big Data in Earth Observation" in 2017 and the IEEE Geoscience and Remote Sensing Society Early Career Award in 2018. She is a senior member of IEEE since 2016.

REFERENCES

[1] Manolis Koubarakis, Konstantina Bereta, George Papadakis, Dimitrianos Savva, and George Stamoulis. 2017. Big, Linked Geospatial Data and Its Applications in Earth Observation. *IEEE Internet Computing* July/August (2017), 87–91.

[2] Manolis Koubarakis, Kostis Kyzirakos, Charalampos Nikolaou, George Garbis, Konstantina Bereta, Roi Doganinad Stella Giannakopoulou, Panayiotis Smeros, Dimitrianos Savva, George Stamoulis, Giannis Vlachopoulos, Stefan Manegold, Charalampos Kontoes, Themistocles Herekakis, Ioannis Papoutsis, and Dimitrios Michail. 2016. Managing Big, Linked, and Open Earth-Observation Data: Using the TELEIOS/LEO software stack. *IEEE Geoscience and Remote Sensing Magazine* 4, 3 (2016), 23–37.