Databases and Information Systems in the AI Era: Contributions from ADBIS, TPDL and EDA 2020 Workshops and Doctoral Consortium

Ladjel Bellatreche[1(✉)], Fadila Bentayeb[2], Mária Bieliková[3], Omar Boussaid[2], Barbara Catania[4], Paolo Ceravolo[5], Elena Demidova[6], Mirian Halfeld Ferrari[7], Maria Teresa Gomez Lopez[8], Carmem S. Hara[9], Slavica Kordić[10], Ivan Luković[10], Andrea Mannocci[11], Paolo Manghi[12], Francesco Osborne[13], Christos Papatheodorou[14], Sonja Ristić[10], Dimitris Sacharidis[15], Oscar Romero[16], Angelo A. Salatino[13], Guilaine Talens[17], Maurice van Keulen[18], Thanasis Vergoulis[19], and Maja Zumer[20]

[1] LIAS/ISAE-ENSMA, Poitiers, France
bellatreche@ensma.fr
[2] Université de Lyon, Lyon 2, ERIC EA 3083, Lyon, France
[3] Slovak University of Technology in Bratislava, Bratislava, Slovakia
[4] University of Genoa, Genoa, Italy
[5] Universitá degli Studi di Milano, Milan, Italy
[6] L3S Research Center, Hannover, Germany
[7] Université d'Orléans, INSA CVL, LIFO EA, Orléans Cedex 2, France
[8] University of Seville, Seville, Spain
[9] Universidade Federal do Paraná, Curitiba, Brazil
[10] Faculty of Technical Sciences, University of Novi Sad, Novi Sad, Serbia
[11] ISTI-CNR, Pisa, Italy
[12] Institute of Information Science and Technologies, CNR, Rome, Italy
[13] The Open University, Milton Keynes, UK
[14] National and Kapodistrian University of Athens, Athens, Greece
papatheodor@phs.uoa.gr
[15] TU Wien, Vienna, Austria
[16] Universitat Politècnica de Catalunya, Barcelona, Spain
[17] Université de Lyon, Jean Moulin, iaelyon, Magellan, Lyon, France
[18] University of Twente, Enschede, The Netherlands
[19] IMSI, Athena Research Center, Athens, Greece
[20] University of Ljubljana, Ljubljana, Slovenia

Abstract. Research on database and information technologies has been rapidly evolving over the last couple of years. This evolution was lead by three major forces: Big Data, AI and Connected World that open the door to innovative research directions and challenges, yet exploiting

four main areas: (i) computational and storage resource modeling and organization; (ii) new programming models, (iii) processing power and (iv) new applications that emerge related to health, environment, education, Cultural Heritage, Banking, etc. The 24th East-European Conference on Advances in Databases and Information Systems (ADBIS 2020), the 24th International Conference on Theory and Practice of Digital Libraries (TPDL 2020) and the 16th Workshop on Business Intelligence and Big Data (EDA 2020), held during August 25–27, 2020, at Lyon, France, and associated satellite events aimed at covering some emerging issues related to database and information system research in these areas. The aim of this paper is to present such events, their motivations, and topics of interest, as well as briefly outline the papers selected for presentations. The selected papers will then be included in the remainder of this volume.

1 Introduction

The East-European Conference on Advances in Databases and Information Systems (ADBIS) aims at providing a forum where researchers and practitioners in the fields of databases and information systems can interact, exchange ideas and disseminate their accomplishments and visions. Inaugurated 24 years ago, ADBIS originally included communities from Central and Eastern Europe, however, throughout its lifetime it has spread and grown to include participants from many other countries throughout the world. The ADBIS conferences provide an international platform for the presentation of research on database theory, development of advanced DBMS technologies, and their advanced applications. The ADBIS series of conferences aims at providing a forum for the presentation and dissemination of research on database theory, development of advanced DBMS technologies, and their advanced applications. ADBIS 2020 in Lyon continues after St. Petersburg (1997), Poznan (1998), Maribor (1999), Prague (2000), Vilnius (2001), Bratislava (2002), Dresden (2003), Budapest (2004), Tallinn (2005), Thessaloniki (2006), Varna (20007), Pori (2008), Riga (2009), Novi Sad (2010), Vienna (2011), Poznan (2012), Genoa (2013), Ohrid (2014), Poitiers (2015), Prague (2016), Nicosia (2017), Budapest (2018) and Bled (2019).

ADBIS 2020 is coupled with TPDL Conference and EDA Workshop. This year, ADBIS, TPDL, and EDA 2020 attract six workshop proposals and Doctoral Consortium.

- The 1st Workshop on Intelligent Data - From Data to Knowledge (DOING 2020), organized by Mirian Halfeld Ferrari (Université d'Orléans, INSA CVL, LIFO EA, France) and Carmem S. Hara (Universidade Federal do Paraná, Curitiba, Brazil).
- The 2nd Workshop on Modern Approaches in Data Engineering and Information System Design (MADEISD 2020), organized by Ivan Luković, Slavica Kordić, and Sonja Ristić (all from University of Novi Sad, Faculty of Technical Sciences, Serbia).

- The 1st Workshop on Scientific Knowledge Graphs (SKG 2020), organized by Andrea Mannocci (ISTI-CNR, Pisa, Italy), Francesco Osborne (The Open University, Milton Keynes, UK) and Angelo A. Salatino (The Open University, Milton Keynes, UK).
- The 2nd Workshop of BI & Big Data Applications (BBIGAP 2020), organized by Fadila Bentayeb and Omar Boussaid (University of Lyon 2, France).
- The Tenth IFIP 2.6 - International Symposium on Data-Driven Process Discovery and Analysis (SIMPDA 2020), organized by Paolo Ceravolo (Universitá degli Studi di Milano, Italy), Maurice van Keulen (University of Twente, The Netherlands) and Maria Teresa Gomez Lopez (University of Seville, Spain).
- The 1st International Workshop on Assessing Impact and Merit in Science (AIMinScience 2020), organized by Paolo Manghi (Institute of Information Science and Technologies, CNR, Italy), Dimitris Sacharidis (TU Wien, Austria), and Thanasis Vergoulis (Athena Research Center, Greece).
- The ADBIS, TPDL & EDA 2020 Doctoral Consortium, organized by Barbara Catania (University of Genoa, Italy), Elena Demidova (L3S Research Center, Germany), Oscar Romero (Universitat Politècnica de Catalunya, Spain) and Maja Zumer (University of Ljubljana, Slovenia).

The main ADBIS, TPDL, and EDA 2020 conferences as well as each of the satellite events had its own international program committee, whose members served as the reviewers of papers included in this volume. This volume contains papers on the contributions of all workshops and the doctoral consortium of ADBIS, TPDL, and EDA 2020. In the following, for each event, we present its main motivations and topics of interest and we briefly outline the papers selected for presentations. The selected papers will then be included in the remainder of this volume. Some acknowledgments from the organizers are finally provided.

2 DOING 2020: The 1st Workshop on Intelligent Data - From Data to Knowledge

Description. The 1st Workshop on Intelligent Data - From Data to Knowledge (DOING 2020), organized by Mirian Halfeld Ferrari (Université d'Orléans, INSA CVL, LIFO EA, France) and Carmem S. Hara (Universidade Federal do Paraná, Curitiba, Brazil).

Texts are important sources of information and communication in diverse domains. The intelligent, efficient and secure use of this information requires, in most cases, the transformation of unstructured textual data into data sets with some structure, and organized according to an appropriate schema that follows the semantics of an application domain. Indeed, solving the problems of modern society requires interdisciplinary research and information cross-referencing, thus surpassing the simple provision of unstructured data. There is a need for representations that are more flexible, subtle, and context-sensitive, which can also be easily accessible via consultation tools and evolve according to these principles. In this context, consultation requires a robust and efficient processing of

queries, which may involve information analysis, quality, consistency, and privacy preservation guarantees. Knowledge bases can be built as these new generation infrastructures which support data science queries on a user-friendly framework and are capable of providing the required machinery for advised decision-making.

DOING Workshop focuses on transforming data into information and then into knowledge. It gathers researchers from Natural Language Processing (NLP), Databases (DB), and Artificial Intelligence (AI). This edition features works in two main areas: (1) information extraction from textual data and its representation on knowledge bases; (2) intelligent methods for handling and maintaining these databases: new forms of requests, including efficient, flexible, and secure analysis mechanisms, adapted to the user, and with quality and privacy preservation guarantees. Overall, the purpose of the workshop is to focus on all aspects concerning modern infrastructures to support these areas, giving particular attention, but not limited, to data on health and environmental domains. DOING received 17 submissions, out of which 8 were accepted as full papers and 1 as a short paper, resulting in an acceptance rate of 50%. Each paper received three reviews from members of the program committee. The accepted papers were allocated in two technical sessions of the workshop program: NLP for information extraction, and Intelligent Data Management. The workshop program also featured an invited keynote talk entitled "Knowledge Graph Completion and Enrichment in OntoSides using Text Mining" by Professor Marie-Christine Rousset, a member of the Laboratoire d'Informatique de Grenoble (LIG). This workshop is an event connected to the working group DOING, involved in the French networks MADICS and RTR DIAMS. The workshop is the result of the collective effort of a large community, which we gratefully acknowledge. We thank the ADBIS-TPDL-EDA joint conference chairs, who worked hard to support the workshop organization. We are also grateful to the members of the program committee, who did an outstanding job, providing timely and thoughtful reviews. Finally, we are grateful to the authors who submitted their work to DOING 2020.

Selected Papers

The *NLP for information extraction Session* includes 4 papers. The first paper, entitled "Extraction of a Knowledge Graph from French Cultural Heritage Documents" [17], describes a method for extracting entities and relations in French heritage descriptive texts, using tools that are available only for English. It presents results using as input texts on Quebec's cultural heritage. Besides heritage-specific type identification, the paper contributes by showing how tools developed for English can be used for other languages with fewer resources available, such as French.

The second paper, "Natural Language Querying System through Entity Enrichment" [2] focuses on translating natural language queries into database queries. The proposed approach is divided into a domain-dependent preprocessing and domain-independent query generation phases. This separation allows the second step to be applied on any domain, although the paper has been motivated by texts in life sciences applications.

The third paper, "Public Riots in Twitter: Domain-Based Event Filtering during Civil Unrest" [19] considers texts from Twitter to identify violence incidents during public riots. The method is composed of 4 steps: temporal clustering, term extraction, scoring, and evaluation. The method was evaluated contrasting the results for a violent and a non-violent event in Peru.

The last paper of the session, "Classification of Relationship in Argumentation using Graph Convolutional Network" [16], is in the area of argumentation mining, which aims to identify claims and evidences from text. The authors propose to model both words and relationships as nodes in a graph and apply a method based on the graph convolutional network to classify relationships among arguments.

The *Intelligent Data Management Session* includes 5 papers. The session starts with the paper "Recursive Expressions for SPARQL Property Paths" [18]. It proposes `rcfSPARQL` (restricted-context-free SPARQL), an extension to SPARQL with a subset of context-free languages. It is based on recursive expressions, which the authors claim that are more user-friendly than context-free languages for writing queries.

The second paper of the session, titled "Healthcare decision-making over a geographic, socioeconomic, and image data warehouse" [25] tackles the problem of processing healthcare analytical queries that involve geographic, socioeconomic and image data. To this end, it proposes three storage models (jointed, split and normalized) and presents an experimental study that determines their performance on a cluster running Spark extended with similarity predicates.

The third paper of the session, titled "Provenance Mechanism for Objects in Deep Learning" [15] proposes OMProv, a mechanism to keep track of the various objects involved in deep learning workflows, as well as their relationships. Each execution is modeled as a weighted directed acyclic graph, and it helps in understanding the outcome of the process. OMProv has been implemented in OMAI, a deep learning platform for the cloud.

The fourth paper, "Exploiting IoT data crossings for gradual pattern mining through parallel processing" [20], proposes a fuzzy approach to mine patterns in time series provided from multiple sources. The algorithm, called FuzzTX, applies a triangular membership function to cross time-series data sets. To show its applicability, the algorithm has been integrated to a Docker implementation of the OGC SensorThings framework.

The paper, "Cooking related Carbon Footprint Evaluation and Optimization" [1] closes the Intelligent Data Management Session. It concerns the carbon footprint of cooking, based on the location of the cooker and the ingredients in a recipe. The authors propose the CaRbon fOotprint reciPe oPtimizER (CROPPER), which takes as input a desired carbon footprint and a money threshold, and generates as output an updated recipe with substitutions of the origin and/or type of its ingredients.

3 MADEISD 2020: The 2nd Workshop on Modern Approaches in Data Engineering and Information System Design

Description. The 2nd Workshop on Modern Approaches in Data Engineering and Information System Design (MADEISD 2020), organized by Ivan Luković, Slavica Kordić and Sonja Ristić (all from University of Novi Sad, Faculty of Technical Sciences, Serbia).

For decades, there is an open issue how to support information management process so as to produce useful knowledge and tangible business values from data being collected. Nowadays, we have a huge selection of various technologies, tools, and methods in data engineering as a discipline that helps in a support of the whole data life cycle in organization systems, as well as in information system design that supports the software process in data engineering. Despite that, one of the hot issues in practice is still how to effectively transform large amounts of daily collected operational data into the useful knowledge from the perspective of declared company goals, and how to set up the information design process aimed at production of effective software services.

The main goal of the Modern Approaches in Data Engineering and Information System Design (MADEISD) workshop is to address open questions and real potentials for various applications of modern approaches and technologies in data engineering and information system design so as to develop and implement effective software services in a support of information management in various organization systems. Intention was to address interdisciplinary character of a set of theories, methodologies, processes, architectures, and technologies in disciplines such as Data Engineering, Information System Design, Big Data, NoSQL Systems, and Model Driven Approaches in a development of effective software services. In this issue, from 9 submissions, after a rigorous selection process, we accepted 4 papers for publication at ADBIS 2020.

Selected Papers

This edition of MADEISD workshop includes four papers.

The authors of the paper "CrEx-Wisdom Framework for fusion of crowd and experts in crowd voting environment - machine learning approach" [14] address the problem of integration of experts domain knowledge with "Wisdom of crowds" by proposing machine learning based framework that enables ranking and selection of alternatives, as well as quantification of quality of crowd votes. The framework proposed by the authors enables weighting of crowd votes with respect to expert knowledge and procedures for modeling trade-off between crowd and experts satisfaction with final decisions based on ranking or selection.

In the paper "Temporal network analytics for fraud detection in the banking sector" [12], the authors present a new methodology in temporal networks for fraud detection in the banking sector. While, standard approaches of fraudulence monitoring mainly have the focus on the individual client data, the authors' approach concentrate on the hidden data produced by the network of a transaction database. The methodology is based on a cycle detection method with the help

of which important patterns can be identified as shown by the test on real data. Proposed solution is integrated into a financial fraud system of a bank.

One of the most common imaging methods for diagnosing an abdominal aortic aneurysm, and an endoleak detection is computed tomography angiography. In the paper "Abdominal Aortic Aneurysm segmentation from contrast-enhanced computed tomography angiography using deep convolutional networks" [10], the authors address the problem of aorta and thrombus semantic segmentation, what is a mandatory step to estimate aortic aneurysm diameter. In the presented research, the three end-to-end convolutional neural networks were trained and evaluated. Finally, the authors proposed an ensemble of deep neural networks with underlying U-Net, ResNet, and VBNet frameworks, and show a possibility to outperform state-of-the-art methods by 3% on the Dice metric without any additional post-processing steps.

One of the latest developments made by publishing companies is introducing mixed and augmented reality to their printed media, e.g. to produce augmented books. An important computer vision problem they are facing with is classification of book pages from video frames. In their paper "Automated classifier development process for recognizing book pages from video frames" [6], the authors address the problem by proposing an automated classifier development process that allows training classification models that run real-time, with high usability, on low-end mobile devices and achieve average accuracy of 88.95% on an in-house developed test set consisting of over 20 000 frames from real videos of 5 books for children.

4 SKG 2020: The 1st Workshop on Scientific Knowledge Graphs

Description. The 1st Workshop on Scientific Knowledge Graphs (SKG 2020), organized by Andrea Mannocci (ISTI-CNR, Pisa, Italy), Francesco Osborne (The Open University, Milton Keynes, UK) and Angelo A. Salatino (The Open University, Milton Keynes, UK).

In the last decade, we experienced a strong need for a flexible, context-sensitive, fine-grained, and machine-actionable representation of scholarly knowledge and corresponding infrastructures for knowledge curation, publishing, and processing. These technical infrastructures are becoming increasingly popular in representing scholarly knowledge as structured, interlinked, and semantically rich scientific knowledge graphs. Knowledge graphs are large networks of entities and relationships, usually expressed in W3C standards such as OWL and RDF. Scientific knowledge graphs focus on the scholarly domain and describe the actors (e.g., authors, organizations), the documents (e.g., publications, patents), and the research knowledge (e.g., research topics, tasks, technologies) in this space as well as their reciprocal relationships. These resources provide substantial benefits to researchers, companies, and policymakers by powering several data-driven services for navigating, analyzing, and making sense of research dynamics.

Some examples include Microsoft Academic Graph (MAG), AMiner, Open Academic Graph, ScholarlyData.org, PID Graph, Open Research Knowledge Graph, OpenCitations, and the OpenAIRE research graph. Current challenges in this area include: i) the design of ontologies able to conceptualise scholarly knowledge, ii) (semi-)automatic extraction of entities and concepts, integration of information from heterogeneous sources, identification of duplicates, finding connections between entities, and iii) the development of new services using this data, that allow exploring this information, measuring research impact and accelerating science.

The 1st Workshop on Scientific Knowledge Graphs (SKG 2020) is a forum for researchers and practitioners from different fields (including, but not limited to, Digital Libraries, Information Extraction, Machine Learning, Semantic Web, Knowledge Engineering, Natural Language Processing, Scholarly Communication, and Bibliometrics) in order to explore innovative solutions and ideas for the production and consumption of scientific knowledge graphs. The scientific program of SKG consists of five papers: three full papers and two short papers, out of 10 submissions, which corresponds to an acceptance rate of 50%. The workshop received submissions from authors of 8 countries in four continents (Europe, Asia, America, Australia). In this edition, three contributions are centered around acquisition, integration and enhancement of scientific knowledge graphs. One contribution covers interoperability between science graphs and another contribution describes a new knowledge organisation system to structure the information within SKGs. Crucially, ontologies are at the core of all submission highlighting the importance of their role in this endeavour.

Selected Papers

The first paper, "Dingo: an ontology for projects and grants linked data" [8], the authors present DINGO (Data INtegration for Grants Ontology), an ontology that provides a machine-readable extensible framework to model data about projects, funding, actors, and funding policies in the research landscape. DINGO is designed to yield high modelling power and elasticity to cope with the wide variety in funding, research and policy practices, which makes it applicable also to other areas besides research where funding is a crucial aspect.

The second paper, "Open science graphs must interoperate!" [5] deals with the major drivers for interoperability of Open Science Graphs (OSGs), Scientific Knowledge Graphs whose represented information may span across entities, such as research artefacts and items of their content, research organisations, researchers, services, projects, funders, and whose intent is to improve the overall FAIRness of science and support stakeholder needs, such as discovery, reuse, reproducibility, statistics, trends, monitoring, validation, and assessment. Despite being valuable individually, OSGs would greatly benefit from information exchange across their collections and, therefore, reuse and exploit the data aggregation and added value that characterise each one of them, decentralising the effort and capitalising on synergies. This work describes the critical motivations for *i)* the definition of a classification for OSGs to compare their features, identify commonalities and differences, and added value and for *ii)* the definition

of an Interoperability Framework, consisting of an information model and APIs that enable a seamless exchange of information across graphs.

The third paper, "WikiCSSH: Extracting Computer Science Subject Headings from Wikipedia" [13] focuses on domain-specific classification schemas (or subject heading vocabularies) which are used to identify, classify, and disambiguate concepts that occur in scholarly articles. Specifically, the authors introduce the Wikipedia-based Computer Science subject headings (WikiCSSH), a large-scale, hierarchically-organised subject heading vocabulary for the domain of Computer Science. It was created by developing, applying, and evaluating a human-in-the-loop workflow that first extracts an initial category tree from crowd-sourced Wikipedia data, and then combines community detection, machine learning, and hand-crafted heuristics or rules to prune the initial tree. WikiCSSH is able to distinguish between coarse-grained and fine-grained CS concepts.

The fourth paper, "Integrating Knowledge Graphs for Analysing Academia and Industry Dynamics" [3] concentrates on knowledge flows between academia and industry. Understanding their mutual influence is a critical task for researchers, governments, funding bodies, investors, and companies. To this end, the authors introduce the Academia/Industry DynAmics (AIDA) Knowledge Graph, which characterises 14M papers and 8M patents according to the research topics drawn from the Computer Science Ontology. 4M papers and 5M patents are also classified according to the type of the author's affiliations (academy, industry, or collaborative) and 66 industrial sectors (e.g., automotive, financial, energy, electronics) obtained from DBpedia. AIDA was automatically generated by integrating different bibliographic corpora, such as Microsoft Academic Graph, Dimensions, English DBpedia, the Computer Science Ontology, and the Global Research Identifier Database.

The last paper, "A Philological Perspective on Meta-Scientific Knowledge Graphs" [30] discusses knowledge graphs and networks on the scientific process from a philological point of view. He argues that all smallest entities or lower-level constituents of science and the scientific text shall be identifiable and contain information on their use throughout all (con)texts, at the same time linking the published version to its parent nodes (e.g. collections, full data sets), and allowing for enquiry through the metadata. Knowledge graphs would then expand the researchers' contributions on a manuscript, as we are reaching conclusions building on the analyses, transcriptions, and interpretations of other scholars.

5 BBIGAP 2020: 2nd Workshop of BI and Big Data Applications

Description. The 2nd Workshop of BI & Big Data Applications (BBIGAP 2020), organized by Fadila Bentayeb and Omar Boussaid (University of Lyon 2, France).

BBIGAP focuses on BI and big data applications. Big Data becomes a huge opportunity for computer science research but it also revolutionizes many fields,

including business, social science, social media, medicine, public administration, and so on. In this case, big data requires a revisit of data management and data analysis techniques in fundamental ways at all stages from data acquisition and storage to data transformation, analysis, and interpretation. The types of available data fall into various categories: social data (e.g., Twitter feeds, Facebook likes), data about mobility and geospatial locations (e.g., sensor data collected through mobile phones or satellite images), data collected from government administrative sources and multilingual text datasets, only to name a few. Big data bring us into a new scientific and technological era offering architectures and infrastructure (clouds, Hadoop-like computing, NoSQL databases) that allow better data management and analytics for decision-making. BBIGAP workshop received 7 submissions, out of which 3 were accepted as full papers and 1 as a short paper, resulting in an acceptance rate of 50%. Each paper received three reviews from members of the program committee.

Selected Papers

This edition of BBIGAP 2020 workshop includes four papers.

The first paper, "A Scored Semantic Cache Replacement Strategy for Mobile Cloud Database Systems" [4] considers the problem of determining the best cache entry to replace in the case of a mobile device accessing a cloud database system. The key idea is that, instead of traditional approaches (e.g., LRU, LFU), replacement techniques must take into account metrics such as current battery life, location, and connectivity quality. For this, they proposed a cache replacement method for mobile cloud database systems that utilizes decisional semantic caching. Specifically, they proposed the Lowest Scored Replacement policy (LSR), a method that uses scored metrics that determine cache relevancy and the mobile devices constraints. The objective is to derive query execution plans (QEPs) expressed as tuples (money, time, energy) and then compute the QEP that best suits the user's requirements.

The second paper, "Grid Based Clustering of Waze Data on a Relational Database" [9] investigates the effect of a grid clustering on the performance of spatial queries, using a relational database. The authors proposed to organize spatiotemporal data as a set of relational tables, using a clustering strategy in order to group together spatiotemporal events. The objective is to show the benefits of organizing the spatial data in these clustering structures for answering spatial queries. Given data collected from traffic events, the authors proposed an approach for partitioning a geographic area of interest. They implemented their approach by using data from Waze over a period of one year, in a specific geographic area. The approach also uses spatial index, like R-trees, to speed up the execution of the type of queries analyzed at the experimental results.

The third paper, "Your Age Revealed by Facebook Picture Metadata" [11] proposes to use machine learning algorithms to infer social media (e.g. Facebook) users' age from metadata related to their pictures. They used logistic regression to classify users' ages into four classes. They showed how sensitive the age information of a given target user can be predicted from his/her online pictures. They investigated the feasibility of age inference attacks on Facebook

users from the metadata of pictures they publish. They showed that commenters react differently to younger and older owner pictures. The proposal is validated with experiments on a data set of 8922 random collected pictures.

The last paper, "Enacting Data Science Pipelines for Exploring Graphs: From Libraries to Studios" [29] provides an overview of data science pipeline libraries, IDE, and studios combining classic and artificial intelligence operations to query, process, and explore graphs. Then, data science pipeline environments are introduced and compared. The paper describes these environments and the design principles that they promote for enacting data science pipelines intended to query, process, and explore data collections and particularly graphs. An example is presented to illustrate how to express graph data science pipelines, that converts data into Data Frame representation and then compute graph metrics.

6 SIMPDA 2020: Tenth IFIP 2.6 - International Symposium on Data-Driven Process Discovery and Analysis

Description. The Tenth IFIP 2.6 - International Symposium on Data-Driven Process Discovery and Analysis (SIMPDA 2020), organized by Paolo Ceravolo (Universitá degli Studi di Milano, Italy), Maurice van Keulen (University of Twente, The Netherlands) and Maria Teresa Gomez Lopez (University of Seville, Spain).

With the increasing automation of business processes, growing amounts of process data become available. This opens new research opportunities for business process data analysis, mining, and modeling. The aim of the IFIP 2.6 - International Symposium on Data-Driven Process Discovery and Analysis is to offer a forum where researchers from different communities and the industry can share their insight into this hot new field.

Our thanks go to the authors, to the program committee, and to who participated in the organization or the promotion of this event. We are very grateful to the Universitá degli Studi di Milano, the University of Seville, the University of Twente, and the IFIP, for supporting this event.

Selected Papers

We selected two papers. In the first paper, "Towards the detection of promising processes by analysing the relational data" [24], the authors cope with the problem of converting relational data into processable event logs observing that multiple views can be obtained from the same relational model fostering different business process analytics.

In the last paper, "Analysis of language inspired trace representation for anomaly detection" [28], the authors develop a comparative study about approaches using vector space modeling for trace profiling. Their comparison can guide the appropriate trace profiling choice for all methods working at the intersection of Process Mining and Machine Learning.

7 AIMinScience 2020: 1st International Workshop on Assessing Impact and Merit in Science

Description. The 1st International Workshop on Assessing Impact and Merit in Science (AIMinScience 2020), organized by Paolo Manghi (Institute of Information Science and Technologies, CNR, Italy), Dimitris Sacharidis (TU Wien, Austria) and Thanasis Vergoulis (Athena Research Center, Greece).

The first edition of the International Workshop on Assessing Impact and Merit in Science (AIMinScience 2020) was held in conjunction with the 24th International Conference on Theory and Practice of Digital Libraries (TPDL 2020). We have compiled these proceedings containing the papers selected for presentation. In the last decades, the growth rate of scientific articles and related research objects (e.g., data sets, software packages) has been increasing, a trend that is expected to continue. This is not only due to the increase in the number of researchers worldwide, but also due to the growing competition that pressures them to continuously produce publishable results, a trend widely known as "publish or perish". This trend has also been correlated with a significant drop in the average quality of published research. In this context, reliable and comprehensive metrics and indicators of the scientific impact and merit of publications, data sets, research institutions, individual researchers, and other relevant entities are now more valuable than ever. Scientific impact refers to the attention a research work receives inside its respective and related disciplines, the social/mass media etc. Scientific merit, on the other hand, is relevant to quality aspects of the work (e.g., its novelty, reproducibility, FAIR-ness, readability). It is evident that any impact or merit metrics and indicators rely on the availability of publication-related (meta)data (e.g., abstracts, citations) which, until recently, were restricted inside the data silos of publishers and research institutions. However today, due to the growing popularity of Open Science initiatives, a large amount of useful science-related data sets have been made openly available, paving the way for more sophisticated scientific impact and merit indicators (and, consequently, more precise research assessment).

Research assessment attracts the interest of researchers and professionals from diverse disciplines. For example, librarians have been working on scientometrics for many decades, while the problems regarding the management and processing of large amounts of scholarly data for research assessment applications has attracted the attention of data scientists recently. The workshop aimed to bring together professionals in academia and industry with such diverse backgrounds being interested in the aforementioned topics and to encourage their interaction. This was facilitated by the fact that this year, TPDL was co-located and co-organized with ADBIS and EDA.

AIMinScience 2020 accepted for presentation 2 full papers and 3 short papers. The program of the workshop also included 3 invited talks and one special session that presented the results of a hackathon. We would like to thank the authors for publishing and presenting their papers, the hackathon participants for their efforts, and our keynote speakers for their talk. We would like to thank the program committee for reviewing the submitted papers and providing their pro-

fessional evaluation. We hope that these proceedings will inspire new research ideas and that you will enjoy reading them.

Keynote Presentations

Predicting the future evolution of scientific output, by Prof. Yannis Manolopoulos.

In the past decade various efforts have been made to quantify scientific impact and in particular identify the mechanisms in play that influence its future evolution. The first step in this process is the identification of what constitutes scholarly impact and how it is measured. In this direction, various approaches focus on future citation count or h-index prediction, either at author or publication level, on fitting the distribution of citation accumulation or accurately identifying award winners, upcoming hot topics in research or academic rising stars. A plethora of different features have been contemplated as possible influential factors in this process and assorted machine-learning methodologies have been adopted to ensure timely and accurate estimations. In the present work, we provide an overview of the challenges rising in the field and a taxonomy of the existing approaches to identify the open issues that are yet to be addressed.

Scientific careers: evolution, interdisciplinarity, gender, and the chaperone effect, by Prof. Roberta Sinatra.

The unprecedented availability of large scale datasets about scholarly output has advanced quantitatively our understanding of how science progresses. In this talk we present a series of findings from the analysis and modelling of large-scale datasets of publications and of scientific careers. We focus on individual scientific careers and tackle the following questions: How does impact evolve in a career? What is the role of gender and of scientific chaperones in dropout and achieving high impact? How interdisciplinary is our recognition system? We show that impact, as measured by influential publications, is distributed randomly within a scientist's sequence of publications, and formulate a stochastic model that uncouples the effects of productivity, individual ability, and luck in scientific careers. We show the role of chaperones in achieving high scientific impact and we study the relation between interdisciplinarity and scientific recognitions. Taken together, we contribute to the understanding of the principles governing the emergence of scientific impact.

Beyond the impact factor: possibilities of scientometrics to understand science and society, by Dr. Rodrigo Costa.

Scientometrics have quite often been related, if not equated, with research evaluation and academic rankings. Multiple debates have emerged about the meaning of citations, the limitations of the Journal Impact Factor or the validity of the h-index as evaluative tools of researchers and research organizations. This strong focus on evaluation may have sometimes concealed other values and uses of scientometric tools regarding research management, and more broadly to study and understand science-society relationships. The main aim of this presentation is to propose and discuss some "non-conventional" uses of scientometric approaches, such as the study of the workforce composition of research organizations, the

tracking of the mobility of researchers across national boundaries, or the interactions between non-academic actors with scholarly objects via social media and altmetrics. These examples are meant to illustrate how the analytical power of scientometric indicators can expand the traditional notions of impact and success.

Selected Papers

The first paper, "Exploring citation networks with hybrid tree pattern queries" [31], proposes to use hybrid query patterns to query citation networks. These allow for both edge-to-edge and edge-to-path mappings between the query pattern and the graph, thus being able to extract both direct and indirect relationships. To efficiently evaluate hybrid pattern queries on citation graphs, a pattern matching algorithm that exploits graph simulation to prune nodes that do not appear in the final answer is applied. The obtained results on citation networks show that the proposed method not only allows for more expressive queries but is also efficient and scalable.

The second paper, "Artsim: Improved estimation of current impact for recent articles" [7] focuses on citation-based measures that try to estimate the popularity (current impact) of a scientific article. The authors identify that the state-of-the-art methods calculate estimates of popularity based on paper citation data. However, with respect to recent publications, only limited data of this type are available, rendering these measures prone to inaccuracies. Based on this finding, the authors present ArtSim, an approach that exploits article similarity, calculated using scholarly knowledge graphs, to better estimate paper popularity for recently published papers. This approach is designed to be applied on top of existing popularity measures, to improve their accuracy. To evaluate its efficiency and effectiveness in terms of improving their popularity estimates, ArtSim is applied on top of four well-known popularity measures.

The third paper, "Link prediction in bibliographic networks" [21] deals with an important problem related to the analysis of bibliographic networks to understanding the process of scientific publications. It should be noticed that a bibliographic network can be studied using the framework of Heterogeneous Information Networks (HINs). The authors compare two different algorithms for link prediction in HINs on an instance of a bibliographic network. These two algorithms represent two distinct categories: algorithms that use path-related features of the graph and algorithms that use node embeddings. The obtained results show that the path-based algorithms achieve significantly better performance on bibliographic networks.

The fourth paper, "Open science observatory: Monitoring open science in Europe" [22] focuses on monitoring and evaluating Open Science (OS) practices and research output in a principled and continuous way. These processes are recognized as one of the necessary steps towards its wider adoption. This paper presents the Open Science Observatory, a prototype online platform that combines data gathered from OpenAIRE e-Infrastructure and other public data sources and informs users via rich visualizations on different OS indicators in Europe.

The last paper, "Skyline-based university rankings" [27] proposes a novel university ranking method based on the skyline operator, which is used on multi-dimensional objects to extract the non-dominated (i.e."prevailing") ones. This method is characterized by several advantages, such as it is transparent, reproducible, without any arbitrarily selected parameters, based on the research output of universities only and not on publicly not traceable or random questionnaires. The proposed method does not provide meaningless absolute rankings but rather it ranks universities categorized in equivalence classes. This method is evaluated using data extracted from Microsoft Academic.

8 DC: The ADBIS, TPDL & EDA 2020 Doctoral Consortium

Description. The ADBIS, TPDL & EDA 2020 Doctoral Consortium has been organized by Barbara Catania (University of Genoa, Italy), Elena Demidova (L3S Research Center, Germany), Oscar Romero (Universitat Politècnica de Catalunya, Spain) and Maja Zumer (University of Ljubljana, Slovenia).

The ADBIS, TPDL & EDA 2020 DC was a forum where Ph.D. students from the database and digital library communities had a chance to present their research ideas. They could gain inspiration and receive feedback from their peers and senior researchers, and to tie cooperation bounds. The DC papers aimed at describing the current status of the thesis research. The DC Committee accepted five presentations, two of which were included in the satellite events proceedings. The topics discussed at the DC included data management, data analysis, social aspects of information technologies, and digitisation of cultural heritage.

Selected Papers

The topics of the two accepted papers concern data analysis and context handling in data management, respectively. In particular, the PhD research described in [23] deals with the definition and the evaluation of a Deep Neural Network (CR-DNN) model to help air traffic controllers to detect situations in which two or more airplanes are less than a minimum distance apart on their trajectory and decide what actions pilots have to apply on the fly. The model learns the best possible action(s) to solve aircraft conflicts based on past decisions and examples.

The research described in [26] deals with Data Quality Management (DQM). Data Quality naturally depends on the application context and usage needs. Despite its recognized importance, the literature lacks of proposals for context definition, specification, and usage within major DQM tasks. Starting from this consideration, the aim of the proposed research is to model and exploit contexts at each phase of the DQM process, providing a proof of concept in the domain of Digital Government.

9 Conclusion

We hope readers will find the content of this volume interesting and will be inspired to look further into the challenges that are still ahead for the design of advanced databases and information systems, with a special reference to Big

Data, AI and Connected World. We are really sure that this volume content will stimulate new ideas for further research and developments by both the scientific and industrial communities.

ADBIS, TPDL & EDA 2020 workshops and Doctoral Consortium organizers would like to express their thanks to everyone who contributed to the volume content. We thank the authors, who submitted papers. Special thanks go to the Program Committee members as well as to the external reviewers of the main conferences and of each satellite event, for their support in evaluating the submitted papers, providing comprehensive, critical, and constructive comments, and ensuring the quality of the scientific program and of this volume.

References

1. Alvarez de Toledo, D., D'Orazio, L., Andres, F., Leite, M.: Cooking related carbon footprint evaluation and optimisation. In: Bellatreche, L., et al. (eds.) ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium, CCIS 1260, pp. 122–128. Springer, Cham (2020)
2. Amavi, J., Halfeld-Ferrari, M., Hiot, N.: Natural language querying system through entity enrichment. In: Bellatreche, L., et al. (eds.) ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium, CCIS 1260, pp. 36–48. Springer, Cham (2020)
3. Angioni, S., Salatino, A., Osborne, F., Recupero, D.R., Enrico Motta, E.: Integrating knowledge graphs for analysing academia and industry dynamics. In: Bellatreche, L., et al. (eds.) ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium, CCIS 1260, pp. 219–225. Springer, Cham (2020)
4. Arani, Z., Chapman, D., Wang, C., Gruenwald, L., D'orazio, L., Basiuk, T.: A scored semantic cache replacement strategy for mobile cloud database systems. In: Bellatreche, L., et al. (eds.) ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium, CCIS 1260, pp. 237–248. Springer, Cham (2020)
5. Aryani, A., Fenner, M., Manghi, P., Mannocci, A., Stocker, M.: Open science graphs must interoperate! In: Bellatreche, L., et al. (eds.) ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium, CCIS 1260, pp. 195–206. Springer, Cham (2020)
6. Brzeski, A., et al.: Automated classifier development process for recognizing book pages from video frames. In: Bellatreche, L., et al. (eds.) ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium, CCIS 1260, pp. 169–179. Springer, Cham (2020)
7. Chatzopoulos, S., Vergoulis, T., Kanellos, I., Dalamagas, T., Tryfonopoulos, C.: Artsim: improved estimation of current impact for recent articles. In: Bellatreche, L., et al. (eds.) ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium, CCIS 1260, pp. 323–334. Springer, Cham (2020)
8. Chialva, D., Mugabushaka, A.M.: Dingo: an ontology for projects and grants linked data. In: Bellatreche, L., et al. (eds.) ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium, CCIS 1260, pp. 183–194. Springer, Cham (2020)
9. Duarte, M.M.G., Schroeder, R., Hara, C.S.: Grid based clustering of waze data on a relational database. In: Bellatreche, L., et al. (eds.) ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium, CCIS 1260, pp. 249–258. Springer, Cham (2020)

10. Dziubich, T., Białas, P., Znaniecki, L., Halman, J., Brzeziński, J.: Abdominal aortic aneurysm segmentation from contrast-enhanced computed tomography angiography using deep convolutional networks. In: Bellatreche, L., et al. (eds.) ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium, CCIS 1260, pp. 158–168. Springer, Cham (2020)
11. Eidizadehakhcheloo, S., Pijani, B.A., Imine, A., Rusinowitch, M.: Your age revealed by Facebook picture metadata. In: Bellatreche, L., et al. (eds.) ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium, CCIS 1260, pp. 259–270. Springer, Cham (2020)
12. Hajdu, L., Krész, M.: Temporal network analytics for fraud detection in the banking sector. In: Bellatreche, L., et al. (eds.) ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium, CCIS 1260, pp. 145–157. Springer, Cham (2020)
13. Han, K., Yang, P., Mishra, S., Diesner, J.: Wikicssh: extracting computer science subject headings from Wikipedia. In: Bellatreche, L., et al. (eds.) ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium, CCIS 1260, pp. 207–218. Springer, Cham (2020)
14. Kovacevic, A., Vukicevic, M., Radovanovic, S., Delibasic, B.: Crex-wisdom framework for fusion of crowd and experts in crowd voting environment - machine learning approach. In: Bellatreche, L., et al. (eds.) ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium, CCIS 1260, pp. 131–144. Springer, Cham (2020)
15. Lin, J., Xie, D.: OMProv: provenance mechanism for objects in deep learning. In: Bellatreche, L., et al. (eds.) ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium, CCIS 1260, pp. 98–109. Springer, Cham (2020)
16. Magalhaes, D., Pozo, A.: Classification of relationship in argumentation using graph convolutional network. In: Bellatreche, L., et al. (eds.) ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium, CCIS 1260, pp. 60–71. Springer, Cham (2020)
17. Marchand, E., Gagnon, M., Zouaq, A.: Extraction of a knowledge graph from French cultural heritage documents. In: In: Bellatreche, L., et al. (eds.) ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium, CCIS 1260, pp. 23–35. Springer, Cham (2020)
18. Medeiros, C.M., Costa, U.S., Grigorev, S.V., Musicante, M.A.: Recursive expressions for SPARQL property paths. In: Bellatreche, L., et al. (eds.) ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium, CCIS 1260, pp. 72–84. Springer, Cham (2020)
19. Oncevay, A., Sobrevilla, M., Alatrista-Salas, H., Melgar, A.: Public riots in Twitter: Domain-based event filtering during civil unrest. In: Bellatreche, L., et al. (eds.) ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium, CCIS 1260, pp. 49–59. Springer, Cham (2020)
20. Owuor, D., Laurent, A., Orero, J.: Exploiting IoT data crossings for gradual pattern mining through parallel processing. In: Bellatreche, L., et al. (eds.) ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium, CCIS 1260, pp. 110–121. Springer, Cham (2020)
21. Chronis, P., Dimitrios Skoutas, S.A., Skiadopoulos, S.: Link prediction in bibliographic networks. In: Bellatreche, L., et al. (eds.) ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium, CCIS 1260, pp. 335–340. Springer, Cham (2020)

22. Papastefanatos, G., et al.: Open science observatory: monitoring open science in Europe. In: Bellatreche, L., et al. (eds.) ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium, CCIS 1260, pp. 341–346. Springer, Cham (2020)
23. Rahman, M.S.: Supervised machine learning model to help controllers solving aircraft conflicts. In: Bellatreche, L., et al. (eds.) ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium, CCIS 1260, pp. 355–361. Springer, Cham (2020)
24. Ramos-Gutiérrez, B., Parody, L., López, M.T.G.: Towards the detection of promising processes by analysing the relational data. In: Bellatreche, L., et al. (eds.) ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium, CCIS 1260, pp. 283–295. Springer, Cham (2020)
25. Rocha, G.M., Capelo, P.L., Dutra De Aguiar Ciferri, C.: Healthcare decision-making over a geographic, socioeconomic, and image data warehouse. In: Bellatreche, L., et al. (eds.) ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium, CCIS 1260, pp. 85–97. Springer, Cham (2020)
26. Serra, F.: Handling context in data quality management. In: Bellatreche, L., et al. (eds.) ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium, CCIS 1260, pp. 362–367. Springer, Cham (2020)
27. Stoupas, G., Antonis Sidiropoulos, D.K., Manolopoulos, Y.: Skyline-based university rankings. In: Bellatreche, L., et al. (eds.) ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium, CCIS 1260, pp. 347–352. Springer, Cham (2020)
28. Tavares, G.M., Junior, S.B.: Analysis of language inspired trace representation for anomaly detection. In: Bellatreche, L., et al. (eds.) ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium, CCIS 1260, pp. 296–308. Springer, Cham (2020)
29. Vargas-Solar, G., Zechinelli-Martini, J., Espinosa-Oviedo, J.A.: Enacting data science pipelines for exploring graphs: from libraries to studios. In: Bellatreche, L., et al. (eds.) ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium, CCIS 1260, pp. 271–280. Springer, Cham (2020)
30. Weber, T.: A philological perspective on meta-scientific knowledge graphs. In: Bellatreche, L., et al. (eds.) ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium, CCIS 1260, pp. 226–233. Springer, Cham (2020)
31. Wu, X., Theodoratos, D., Skoutas, D., Lan, M.: Exploring citation networks with hybrid tree pattern queries. In: Bellatreche, L., et al. (eds.) ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium, CCIS 1260, pp. 311–322. Springer, Cham (2020)