Preface

This book is inspired by the SYNAT 2013 Workshop held in Warsaw, which was a forum for exchange of experience in the process of building a scientific information platform. It is a consecutive book of the series related to the SYNAT project, and it summarizes the workshop's results as well as captures the current work progress in the project. The previous volumes entitled "Intelligent Tools for Building a Scientific Information Platform" and "Intelligent Tools for Building a Scientific Information Platform: Advanced Architectures and Solutions," have been also published in Springer's Studies in Computational Intelligence.

SYNAT is a program that was initiated in 2010 and has been scheduled for a period of 3 years. It has been focused on meeting the challenges of global digital information revolution, especially in the context of scientific information. The final results of the program encompass two main areas:

- research on consistent strategies for the development of domain-specific systems of repositories that constitute a basis for future-oriented activities in the areas recognized as critical for the national knowledge infrastructure, inter alia, artificial intelligence, knowledge discovery and data mining, information retrieval, and natural language processing,
- development of an integrated ICT platform for operating the entire complex of knowledge resources (i) equipped with a comprehensive set of functionalities, among those a range of novel tools supporting enhanced forms of scholarly communication, (ii) allowing for two-way collaborations on various levels, breaking the barriers between creation and use, (iii) facilitating its interoperability with leading international knowledge infrastructures.

The papers included in this volume cover topics of diverse character, which is reflected in the arrangement of this book. The book consists of the following parts: Challenges, Research, Research Environments, and Implemented Systems. We will now outline the contents of the chapters.

Part I, "Challenges," deals with challenges facing information science and presents trends likely to shape it in the years to come as well as gives an outline of an informational model of open innovation.

- Bruno Jacobfeuerborn and Mieczysław Muraszkiewicz ("Some Challenges and Trends in Information Science") argue that contemporary information science is

a vivid discipline whose development is inspired and driven by many other disciplines such as theory of information, mathematics, computer science, psychology, sociology and social communications, librarianship, museology and archival sciences, linguistics, law, and cognitive sciences. In the chapter the authors briefly take a look at a collection of assorted phenomena, methodologies, and technologies that will have a significant impact on the scope and the ways information science will unfold in the coming years. The chapter provides the authors' understanding of information science as a dynamic discipline that extends its boundaries as new methodologies and technologies come along, as a result of scientific discoveries, engineering achievements, and emergence of new business models. It presents and elaborates on the challenges that constitute a framework within which new trends in information science have started appearing and most likely will shape the face of information science and its applications in the years to come.

- Bruno Jacobfeuerborn ("An Informational Model of Open Innovation") proposes an outline of an informational model of open innovation. This model can be a foundation for establishing an innovation facility by a production or service provider entity. The pillars of the model are: (i) access to and extensive use of up-to-date information, knowledge, and the best practices; (ii) the use of efficient and user friendly knowledge management systems based on advanced semantic tools; (iii) extensive use of ICT business intelligence tools, including Web 2.0 facilities and Web services, and social networks; (iv) a customer experience management system and customers' participation in identifying new products and/or value-added services (prosumerism); (v) close collaboration with academia and relevant nongovernmental organizations; and (vi) access to a collaborative work platform anytime and from anywhere, implemented as a cloud computing facility. It seems that the proposals included in the chapter can contribute to the research within new areas and trends in management, in particular with respect to modeling innovation processes and their evaluation.

Part II, "Research", is devoted to theoretical studies in the areas of artificial intelligence, machine learning, text processing, and knowledge engineering addressing the problems of implementing intelligent tools for building a scientific information platform.

- Marzena Kryszkiewicz and Bartłomiej Jańczak ("Basic Triangle Inequality Approach versus Metric VP-Tree and Projection in Determining Euclidean and Cosine Neighbors") discuss three approaches to efficient determination of nearest neighbors, namely using the triangle inequality when vectors are ordered with respect to their distances to one reference vector, using a metric VP-tree, and using a projection onto a dimension. The techniques are well suited to any distance metrics such as the Euclidean distance, but they cannot be directly used for searching nearest neighbors with respect to the cosine similarity. The authors provide an experimental comparison of the three techniques for determining nearest neighbors with regard to the Euclidean distance and the cosine similarity.

- Karol Draszawka, Julian Szymański and Henryk Krawczyk ("Towards Increasing Density of Relations in Category Graphs") proposes methods for identifying new associations between Wikipedia categories. The first method is based on Bag-of-Words (BOW) representation of Wikipedia articles. Using similarity of the articles belonging to different categories allows to calculate the information about categories' similarity. The second method is based on average scores given to categories while categorizing documents by dedicated score-based classifier.
- Hung Son Nguyen, Michał Meina, and Wojciech Świeboda ("Weight Learning in TRSM-based Information Retrieval") present a novel approach to keyword search in Information Retrieval based on Tolerance Rough Set Model (TRSM). Bag-of-Word representation of each document is extended by additional words that are enclosed into inverted index along with appropriate weights. The extension words are derived from different techniques (e.g., semantic information, word distribution, etc.) that are encapsulated in the model by a tolerance relation. Weight for structural extension are then assigned by unsupervised algorithm.
- Krzysztof Goczyła, Aleksander Waloszek, and Wojciech Waloszek ("An Analysis of Contextual Aspects of Conceptualization: A Case Study and Prospects") present a new approach to development of modularized knowledge bases. The authors argue that modularization should start from the very beginning of modeling, i.e., from the conceptualization stage. To make this feasible, they propose to exploit a context-oriented, semantic approach to modularization. This approach is based on the Structural Interpretation Model (SIM) presented earlier elsewhere. The first part of the chapter presents a contextualized version of the SYNAT ontology developed using the SIM methodology as well as a set of tools needed to enable a knowledge engineer to create, edit, store, and perform reasoning over contextualized ontologies in a flexible and natural way. The second part of the chapter gives a deeper insight into some aspects of using these tools, as well as into ideas underlying their construction. The work on contextualization of knowledge bases led the authors to further theoretical investigation of the hierarchical structure of a knowledge base system. Indeed, in a system of heterogeneous knowledge sources, each source (a knowledge base) can be seen in its own separate context, as being a part of a higher level contextual structure (a metastructure) with its own set of context parameters. The theoretical background of this conception is presented in the third part of the chapter.

Part III, "Research Environments," presents environments aimed at research purposes created within SYNAT. These environments include: Music Discovery and Recommendation System, Web Resource Acquisition System for Building Scientific Information Database, tools for log analysis (LogMiner, FEETS, ODM), Content Analysis System (CoAnSys), System for Fast Text Search and Document Comparison, Chrum—a tool for convenient generation of apache Oozie Workflows and PrOnto—A Local Search Engine for Digital Libraries.

- Bożena Kostek, Piotr Hofmann, Andrzej Kaczmarek, and Paweł Spaleniak ("Creating a Reliable Music Discovery and Recommendation System") discuss problems related to creating a reliable music discovery system. The SYNAT database that contains audio files is used for the purpose of experiments. The files are divided into 22 classes corresponding to music genres with different cardinality. Of utmost importance for a reliable music recommendation system are the assignment of audio files to their appropriate genres and optimum parameterization for music-genre recognition. Hence, the starting point is audio file filtering, which can only be done automatically, but to a limited extent, when based on low-level signal processing features. Therefore, a variety of parameterization techniques are briefly reviewed in the context of their suitability to music retrieval from a large music database. In addition, some significant problems related to choosing an excerpt of audio file for an acoustic analysis and parameterization are pointed out. Then, experiments showing results of searching for songs that bear the greatest resemblance to the song in a given query are presented. In this way a music recommendation system may be created that enables to retrieve songs that are similar to each other in terms of their low-level feature description and genre inclusion. The experiments performed also provide the basis for more general observations and conclusions.
- Tomasz Adamczyk and Piotr Andruszkiewicz ("Web Resource Acquisition System for Building Scientific Information Database") describe the architecture and findings made as an effect of integration of a complex resource acquisition system with a frontend system. Both underlying and frontend systems are mentioned briefly with reference to their root publications. The main accent has been put on data architecture, information processing, user interaction, obtained results, and possible future adaptation of the system.
- Janusz Sosnowski, Piotr Gawkowski, Krzysztof Cabaj, and Marcin Kubacki ("Analyzing Logs of the University Data Repository") claim that identification of execution anomalies is very important for the maintenance and performance refinement of computer systems. For this purpose they use system logs. These logs contain vast amounts of data, hence there is a great demand for techniques targeted at log analysis. The chapter presents authors' experience with monitoring event and performance logs related to data repository operation. Having collected representative data from the monitored systems the authors have developed original algorithms of log analysis and problem predictions that are based on various data mining approaches. These algorithms have been included in the implemented tools: LogMiner, FEETS, and ODM. Practical significance of the developed approaches has been illustrated with some examples of exploring data repository logs. To improve the accuracy of problem diagnostics the authors have developed supplementary log database which can be filled in by system administrators and users.
- Piotr Jan Dendek, Artur Czeczko, Mateusz Fedoryszak, Adam Kawa, Piotr Wendykier, and Łukasz Bolikowski ("Content Analysis of Scientific Articles in Apache Hadoop Ecosystem") describe algorithms currently implemented in CoAnSys (Content Analysis System), which is a research framework for mining

scientific publications using Apache Hadoop. The algorithms include classification, categorization, and citation matching of scientific publications. The size of the input data classifies these algorithms in the range of big data problems, which can be efficiently solved on Hadoop clusters.

- Maciej Wielgosz, Marcin Janiszewski, Marcin Pietroń, Pawel Russek, Ernest Jamro, and Kazimierz Wiatr ("Implementation of a System for Fast Text Search and Document Comparison") present an architecture of a system for fast text search and documents comparison with a main focus on N-gram-based algorithm and its parallel implementation. The algorithm which is one of several computational procedures implemented in the system is used to generate a fingerprint of analyzed documents as a set of hashes which represent the file. The work examines the performance of the system, both in terms of a file comparison quality and a fingerprint generation. Several tests were conducted of N-gram-based algorithm for Intel Xeon E5645, 2.40 GHz which show approximately 8x speedup of multi over single core implementation.
- Piotr Jan Dendek, Artur Czeczko, Mateusz Fedoryszak, Adam Kawa, Piotr Wendykier, and Łukasz Bolikowski ("Chrum: the Tool for Convenient Generation of Apache Oozie Workflows") argue that conducting a research in an efficient, repetitive, evaluable, but also convenient (interms of development) way has always been a challenge. To satisfy these requirements in a long term and simultaneously minimize costs of the software engineering process, one has to follow a certain set of guidelines. The article describes such guidelines based on the research environment called Content Analysis System (CoAnSys) created in the Center for Open Science (CeON). In addition to the best practices for working in the Apache Hadoop environment, the tool for convenient generation of Apache Oozie workflows is presented.
- Janusz Granat, Edward Klimasara, Anna Mościcka, Sylwia Paczuska, and Andrzej Piotr Wierzbicki ("PrOnto: A Local Search Engine for Digital Libraries") describe system PrOnto version 2.0 and results of work on this system in the SYNAT project. After the introduction, the chapter presents briefly the functionality of PrOnto that is a system of personalized search for information and knowledge in large text repositories. Further, the chapter presents elements of the personalized ontological profile of the user, the problem of finding similar concepts in many such profiles, and the issue of finding interesting documents in large text repositories, together with tests of the system and conclusions.

Part IV, "Implemented Systems," showcases production systems that were developed during the course of the SYNAT project.

- Andrzej Czyżewski, Adam Kupryjanow, and Janusz Cichowski ("Further Developments of the Online Sound Restoration System for Digital Library Applications") describe new signal processing algorithms that were introduced to the online service for audio restoration available at the web address: http://www.youarchive.net. Missing or distorted audio samples are estimated using a

specific implementation of the Jannsen interpolation method. The algorithm is based on the autoregressive model (AR) combined with the iterative complementation of signal samples. The chapter is concluded with a presentation of experimental results of application of described algorithmic extensions to the online sound restoration service.

- Adam Dudczak, Michał Dudziński, Cezary Mazurek, and Piotr Smoczyk ("Overview of Virtual Transcription Laboratory Usage Scenarios and Architecture") outline findings from the final stage of development of the Virtual Transcription Laboratory (http://wlt.synat.pcss.pl, VTL) prototype. VTL is a crowdsourcing platform developed to support creation of the searchable representation of historic textual documents from Polish digital libraries. This chapter describes identified usage scenarios and shows how they were implemented in the data model and architecture of the prototype.
- Jakub Koperwas, Łukasz Skonieczny, Marek Kozłowski, Henryk Rybiński, and Wacław Struk ("University Knowledge Base: Two Years of Experience") summarize 2-years development and exploitation of the repository platform built at Warsaw University of Technology for the purpose of gathering University research knowledge. The implementation of the platform in the form of the advanced information system is discussed. New functionalities of the knowledge base are presented.
- Cezary Mazurek, Marcin Mielnicki, Aleksandra Nowak, Krzysztof Sielski, Maciej Stroiński, Marcin Werla, and Jan Węglarz ("CLEPSYDRA Data Aggregation and Enrichment Framework: Design, Implementation and Deployment in the PIONIER Network Digital Libraries Federation") describe the architecture for aggregation, processing, and provisioning of data from heterogeneous scientific information services. The implementation of this architecture was named CLEPSYDRA and was published as an open source project. This chapter contains an overview of the CLEPSYDRA system design and implementation. It also presents the test deployment of CLEPSYDRA for the purpose of the PIONIER Network Digital Libraries Federation, focusing on aspects such as agent-based data aggregation, data normalization, and data enrichment. Finally, the chapter includes several scenarios for the future use of the system in the national and international contexts.

This book could not have been completed without the involvement of many people. We would like to express our high appreciation to all the contributors and to thank the reviewers. We are grateful to Janusz Kacprzyk for encouraging us to prepare this book.

Warsaw, November 2013

Robert Bembenik
Łukasz Skonieczny
Henryk Rybiński
Marzena Kryszkiewicz
Marek Niezgódka

http://www.springer.com/978-3-319-04713-3

Intelligent Tools for Building a Scientific Information Platform: From Research to Implementation

Bembenik, R.; Skonieczny, L.; Rybiński, H.; Kryszkiewicz, M.; Niezgódka, M. (Eds.)

2014, XIII, 290 p. 105 illus., Hardcover
ISBN: 978-3-319-04713-3