Book of Abstracts and Posters

Metadata and

Semantics Research

11th International Research Conference,

MTSR 2017

November 28th – December 1st, 2017

Tallinn, Estonia

Book of Abstracts

Editing and Design

Emmanouel Garoufallou

Working Group

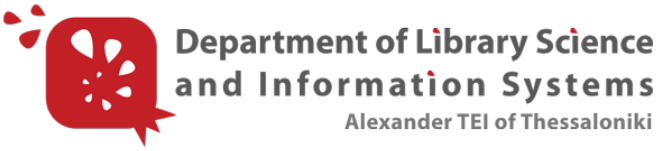
Iro Sotiriadou, Anxhela Dani, Chrysanthi Chatzopoulou, Rania Siatri, Damiana Koutsomiha, Stavroula Antonopoulou, Sirje Virkus

Alexander Technological Educational Institute of Thessaloniki

Co-organized by

Tallinn University

Department of Library Science and Information Systems | Alexander TEI of Thessaloniki



Alexander Technological Educational Institute of Thessaloniki

Awards sponsored by

euroCRIS
Current Research Information Systems

Sponsors

Springer

ReasonableGraph.org
Thinking Ontologies

Cataloging-in-Publication Data

International Conference on Metadata and Semantics Research (11th : 2017 : Tallinn, Estonia)

Metadata and semantics research : 11th Research Conference, MTSR 2017, November 28 – December 1 2017, Tallinn, Estonia : book of abstracts / editing and design Emmanouel Garoufallou ; working group Iro Sotiriadou … [et al.]

p. cm.

Includes index.

ISBN

1. Metadata–Congresses. 2. Semantic web–Congresses. 3. Semantics–Data processing–Congresses. 4. Semantic computing–Congresses. I. Garoufallou, Emmanouel. II. Sotiriadou, Iro.

006’.7’–dc23

Table of Contents

Preface

Since 2005, the International Metadata and Semantics Research Conference (MTSR) has served as a significant venue for the dissemination and sharing of metadata and semantic-driven research and practices. This year, 2017, marked the 11th MTSR, drawing scholars, researchers and practitioners who are investigating and advancing our knowledge on a wide range of metadata and semantic-driven topics. The 11th International Conference on Metadata and Semantics Research (MTSR'17) was held at Tallinn University (Estonia) from November 28th to December 1st, 2017.

Metadata and semantics are integral to any information system and important to the sphere of Web data. Research and development addressing metadata and semantics is crucial to advancing how we effectively discover, use, archive, and repurpose information. In response to this need, researchers are actively examining methods for generating, reusing, and interchanging metadata. Integrated with these developments is research on the application of computational methods, linked data, and data analytics. A growing body of literature also targets conceptual and theoretical designs providing foundational frameworks for metadata and semantic applications. There is no doubt that metadata weaves its way through nearly every aspect of our information ecosystem, and there is great motivation for advancing the current state of understanding in the fields of metadata and semantics. To this end, it is vital that scholars and practitioners convene and share their work.

MTSR 2017 focused on an emerging theme of “Internet of Things (IoT) in Library and Information Science Research” and the practical implementation of ontologies and linked data in various applications. The conference focuses on: Theoretical and foundational principles of metadata, ontologies and information organization; The emergence and application of the Internet of Things (IoT) in libraries and cultural heritage institutions (such as RFID technologies, smart libraries and virtual museums); The applications of Linked Data, Open Data, Big Data, and user-generated metadata; Digital Interconnectedness – the what, why and how of Linked Open Data, and the Semantic Web; Metadata standardization, authority control and interoperability in digital libraries, and research data repositories; Emerging issues in RDF, OWL, SKOS, schema.org, BIBFRAME, metadata and ontology design; Linked data applications for e-books, digital publishing and Content Management Systems (CMSs); Content discovery services, search, information retrieval, and data visualization applications.

MTSR conferences have grown in number of participants and paper submission rates over the last decade, marking it as a leading, international research conference. Continuing in the successful legacy of previous MTSR conferences (MTSR 2005, MTSR 2007, MTSR 2009, MTSR 2010, MTSR 2011, MTSR 2012, MTSR 2013, MTSR 2014, MTSR 2015, and MTSR 2016), MTSR 2017 brought together scholars and practitioners who share a common interest in the interdisciplinary field of metadata, linked data, and ontologies.

The MTSR 2017 program and the following proceedings show a rich diversity of research and practices from metadata and semantically focused tools and technologies, linked data, cross language semantics, ontologies, metadata models, semantic systems, and metadata standards. The general session of the conference included nine papers covering a broad spectrum of topics, proving the interdisciplinary view of metadata. Metadata as a research topic is maturing, and the conference supported the following seven tracks: Digital Libraries, Information Retrieval, Big, Linked, Social, and Open Data; Metadata and Semantics for Cultural Collections and Applications; Track on European and National Projects; Metadata and Semantics for Open Repositories, Research Information Systems and Data Infrastructures; Track on Digital Humanities and Digital Curation; Metadata and Semantics for Agriculture, Food, and Environment; Track on Knowledge IT Artifacts in Professional Communities and Aggregations. Each of these tracks had a rich selection of short and full research papers, in total 22, giving broader diversity to MTSR, and enabling deeper exploration of significant topics.

All the papers underwent a thorough and rigorous peer-review process. The review and selection for this year were highly competitive and only papers containing significant research results, innovative methods, or novel and best practices were accepted for publication. From the general session, only five submissions were accepted as full research papers, representing 33.3% of the total number of submissions, and four as short papers. An additional 13 contributions from tracks covering noteworthy and important results were accepted as full research papers representing 30.2% of the total number of submissions, and nine as short papers, making up 31 contributions for this year's MTSR. The acceptance rate of full research papers for both the general session and tracks was 31% of the total number of submissions.

Tallinn University is the third largest public university in Estonia, focusing primarily on the fields of humanities and the social and natural sciences. In its activities, the university adheres to the following basic values: openness, quality, professionalism, and unity. The study area of information sciences of the School of Digital Technologies is the co-organizer of the MTSR 2017. The School of Digital Technologies aims to integrate the study areas of digital learning ecosystems, information sciences, human-computer interaction, mathematics and didactics of mathematics, and applied informatics in order to develop interdisciplinary competencies related to the digital information and learning environment and information and digital competencies.

This year the MTSR conference was pleased to host one remarkable keynote presentation by Dr. Trond Aalberg, Associate Professor in the Department of Computer Science, Data and Artificial Intelligence Group at NTNU (Norwegian University of Science and Technology). In his presentation "The Path Toward Bibliographic Ontologies and Linked Data," Professor Aalberg shared his extensive experience and insights about the various models of bibliographic data, transformation of existing data, quality issues, reuse, and use of such data in search systems.

We conclude this preface by thanking the many people who contributed their time and efforts to MTSR 2017 and made this year's conference possible. We also thank all the

organizations that supported this conference. We extend a sincere gratitude to members of the Program Committees, both main and special tracks, the Steering Committee, and the Organizing Committees (both general and local), and the conference reviewers who invested their time generously to ensure the timely review of the submitted manuscripts. A special thanks to Program Chair Dr. Getaneh Alemu from Southampton Solent University, UK; to Anxhela Dani and Pirje Jürgens for supporting us throughout this event and to Iro Sotiriadou and Anxhela Dani, who assisted us with the preparation of proceedings; to Iro, Anxhela, Chrysanthi Chatzopoulou, and the rest of the MTSR team, who assisted us with the preparation of the Book of Abstracts, and to Stavroula, Nikoleta, and Vasiliki for their endless support and patience. Our thanks go to all participants of MTSR 2017 for making the event a great success.

Flogita, Tallinn, Thessaloniki
September 2017

Emmanouel Garoufallou
Sirje Virkus
Rania Siatri
Damiana Koutsomiha

Organization

General Chairs

Emmanouel Garoufallou	Alexander Technological Educational Institute (ATEI) of Thessaloniki, Greece
Sirje Virkus	Tallinn University, Estonia

Program Chair

Getaneh Alemu	Southampton Solent University, UK

Special Track Chairs

Ernesto William De Luca	Georg Eckert-Institute – Leibniz-Institute for International Textbook Research, Germany
Paolo Bianchini	Università degli Studi di Torino, Italy
Miguel-Ángel Sicilia	University of Alcalá, Spain
Armando Stellato	University of Rome Tor Vergata, Italy
Juliette Dibie	AgroParisTech and INRA, France
Liliana Ibanescu	AgroParisTech and INRA, France
Michalis Sfakakis	Ionian University, Greece
Lina Bountouri	Ionian University, Greece and EU Publications Office, Luxemburg
Emmanouel Garoufallou	Alexander Technological Educational Institute (ATEI) of Thessaloniki, Greece
R. J. Hartley	Manchester Metropolitan University, UK
Rania Siatri	Alexander Technological Educational Institute (ATEI) of Thessaloniki, Greece
Stavroula Antonopoulou	Perrotis College, Greece
Georgia Zafeiriou	University of Macedonia, Greece
Fabio Sartori	University of Milano-Bicocca, Italy
Angela Locoro	University of Milano-Bicocca, Italy
Arlindo Flavio da Conceição	Federal University of São Paulo (UNIFESP), Brazil

Steering Committee

Juan Manuel Dodero	University of Cádiz, Spain

Emmanouel Garoufallou	Alexander Technological Educational Institute (ATEI) of Thessaloniki, Greece
Nikos Manouselis	AgroKnow, Greece
Fabio Santori	University of Milano-Bicocca, Italy
Miguel-Ángel Sicilia	University of Alcalá, Spain

Organizing Committee

Aile Möldre	Tallinn University, Estonia
Elviine Uverskaja	Tallinn University, Estonia
Silvi Metsar	Tallinn University, Estonia
Sigrid Mandre	Tallinn University, Estonia
Hans Põldoja	Tallinn University, Estonia
Pirje Jürgens	Tallinn University, Estonia
Damiana Koutsomiha	American Farm School, Greece
Anxhela Dani	Alexander Technological Educational Institute (ATEI) of Thessaloniki, Greece
Chrysanthi Chatzopoulou	Alexander Technological Educational Institute (ATEI) of Thessaloniki, Greece
Iro Sotiriadou	American Farm School, Greece
Panorea Gaitanou	Ionian University, Greece
Ioanna Andreou	Hellenic-American Educational Foundation, Greece

Technical Support Staff

Ilias Nitsos	Alexander Technological Educational Institute (ATEI) of Thessaloniki, Greece

Program Committee

Rajendra Akerkar	Western Norway Research Institute, Norway
Arif Altun	Hacettepe University, Turkey
Ioannis N. Athanasiadis	Democritus University of Thrace, Greece
Panos Balatsoukas	University of Manchester, UK
Tomaz Bartol	University of Ljubljana, Slovenia
Ina Bluemel	German National Library of Science and Technology TIBm, Germany
Derek Bousfield	Manchester Metropolitan University, UK
Gerhard Budin	University of Vienna, Austria

Özgü Can	Ege University, Turkey
Caterina Caracciolo	Food and Agriculture Organization (FAO) of the United Nations, Italy
Christian Cechinel	Federal University of Pampa, Brazil
Artem Chebotko	University of Texas - Pan American, USA
Philip Cimiano	Bielefeld University, Germany
Sissi Closs	Karlsruhe University of Applied Sciences, Germany
Ricardo Colomo-Palacios	Universidad Carlos III, Spain
Sally Jo Cunningham	Waikato University, New Zealand
Constantina Costopoulou	Agricultural University of Athens, Greece
Ernesto William De Luca	Georg Eckert-Institute – Leibniz-Institute for International Textbook Research, Germany
Milena Dobreva	University of Malta, Malta
Juan Manuel Dodero	University of Cádiz, Spain
Erdogan Dogdu	TOBB Teknoloji ve Ekonomi University, Turkey
Juan José Escribano Otero	Universidad Europea de Madrid, Spain
Muriel Foulonneau	Tudor Public Research Centre, Luxemburg
Panorea Gaitanou	Ionian University, Greece
Emmanouel Garoufalou	Alexander Technological Educational Institute (ATEI) of Thessaloniki, Greece
Manolis Gergatsoulis	Ionian University, Greece
Jorge Gracia Del Río	Universidad Politécnica de Madrid, Spain
Jane Greenberg	Drexel University, USA
Jill Griffiths	Manchester Metropolitan University, UK
R. J. Hartley	Manchester Metropolitan University, UK
Nikos Houssos	Redlink, Greece
Carlos A. Iglesias	Universidad Politécnica de Madrid, Spain
Frances Johnson	Manchester Metropolitan University, UK
Dimitris Kanellopoulos	University of Patras, Greece
Pinar Karagöz	Middle East Technical University (METU), Turkey
Pythagoras Karampiperis	AgroKnow, Greece
Brian Kelly	CETIS, University of Bolton, UK
Christian Kop	University of Klagenfurt, Austria
Rebecca Koskela	University of New Mexico, USA
Daniela Luzi	National Research Council, Italy
Paolo Manghi	Institute of Information Science and Technologies (ISTI), National Research Council (CNR), Italy

John McCrae	National University of Ireland Galway, Ireland
Xavier Ochoa	Centro de Tecnologías de Información Guayaquil, Ecuador
Mehmet C. Okur	Yaşar University, Turkey
Matteo Palmonari	University of Milano-Bicocca, Italy
Manuel Palomo Duarte	University of Cádiz, Spain
Laura Papaleo	University of Genova, Italy
Christos Papatheodorou	Ionian University, Greece
Marios Poulos	Ionian University, Greece
T. V. Prabhakar	Indian Institute of Technology Kanpur, India
Maria Cláudia Reis Cavalcanti	Military Institute of Engineering, Brazil
Athena Salaba	Kent State University, USA
Salvador Sánchez-Alonso	University of Alcalá, Spain
Fabio Sartori	University of Milano-Bicocca, Italy
Cleo Sgouropoulou	Technological Educational Institute of Athens, Greece
Rania Siatri	Alexander Technological Educational Institute (ATEI) of Thessaloniki, Greece
Miguel-Ángel Sicilia	University of Alcalá, Spain
Armando Stellato	University of Rome Tor Vergata, Italy
Imma Subirats	Food and Agriculture Organization (FAO) of the United Nations, Italy
Shigeo Sugimoto	University of Tsukuba, Japan
Stefaan Ternier	Open University of the Netherlands, Netherlands
Giannis Tsakonas	University of Patras, Greece
Andrea Turbati	University of Rome Tor Vergata, Italy
Fabio Massimo Zanzotto	University of Rome Tor Vergata, Italy
Thomas Zschocke	World Agroforestry Centre (ICRAF), Kenya

Track on Metadata and Semantics for Digital Libraries, Information Retrieval, Big, Linked, Social and Open Data

Special Track Chairs

Emmanouel Garoufallou	Alexander Technological Educational Institute (ATEI) of Thessaloniki, Greece
Rania Siatri	Alexander Technological Educational Institute (ATEI) of Thessaloniki, Greece
Sirje Virkus	Tallinn University, Estonia

Program Committee

Panos Balatsoukas	University of Manchester, UK
Özgü Can	Ege University, Turkey
Sissi Closs	Karlsruhe University of Applied Sciences, Germany
Mike Conway	University of North Carolina at Chapel Hill, USA
Phil Couch	University of Manchester, UK
Milena Dobreva	University of Malta, Malta
Ali Emrouznejad	Aston University, UK
Panorea Gaitanou	Ionian University, Greece
Jane Greenberg	Drexel University, USA
Jill Griffiths	Manchester Metropolitan University, UK
R. J. Hartley	Manchester Metropolitan University, UK
Nikos Korfiatis	University of East Anglia, UK
Rebecca Koskela	University of New Mexico, USA
Valentini Moniarou-Papaconstantinou	Technological Educational Institute of Athens, Greece
Dimitris Rousidis	University of Alcalá, Spain
Athena Salaba	Kent State University, USA
Rania Siatri	Alexander Technological Educational Institute (ATEI) of Thessaloniki, Greece
Miguel-Ángel Sicilia	University of Alcalá, Spain
Christine Urquhart	Aberystwyth University, UK
Evgenia Vassilakaki	Technological Educational Institute of Athens, Greece
Sirje Virkus	Tallinn University, Estonia
Georgia Zafeiriou	University of Macedonia, Greece

Track on Metadata and Semantics for Cultural Collections and Applications

Special Track Chairs

Michalis Sfakakis	Ionian University, Greece
Lina Bountouri	Ionian University, Greece and EU Publications Office, Luxemburg

Program Committee

Trond Aalberg	Norwegian University of Science and Technology (NTNU), Norway
Karin Bredenberg	The National Archives of Sweden, Sweden
Enrico Francesconi	EU Publications Office, Luxembourg, and Consiglio Nazionale delle Ricerche, Firenze, Italy
Manolis Gergatsoulis	Ionian University, Greece
Antoine Isaac	Vrije Universiteit Amsterdam, The Netherlands
Sarantos Kapidakis	Ionian University, Greece
Peter McKinney	National Library of New Zealand, New Zealand
Christos Papatheodorou	Ionian University and Digital Curation Unit, IMIS, Athena RC, Greece
Chrisa Tsinaraki	Joint Research Centre, European Commission, Italy
Andreas Vlachidis	University of South Wales, UK
Katherine Wisser	Simmons College, USA
Maja Žumer	University of Ljubljana, Slovenia

Track on Metadata and Semantics for European and National Projects

Special Track Chairs

Stavroula Antonopoulou	Perrotis College, Greece
R. J. Hartley	Manchester Metropolitan University, UK
Georgia Zafeiriou	University of Macedonia, Greece
Rania Siatri	Alexander Technological Educational Institute (ATEI) of Thessaloniki, Greece

Program Committee

Stavroula Antonopoulou	Perrotis College, Greece
Panos Balatsoukas	University of Manchester, UK
Mike Conway	University of North Carolina at Chapel Hill, USA
Jane Greenberg	Drexel University, USA
R. J. Hartley	Manchester Metropolitan University, UK
Nikos Houssos	RedLink, Greece
Nikos Korfiatis	University of East Anglia, UK
Damiana Koutsomiha	American Farm School, Greece
Paolo Manghi	Institute of Information Science and Technologies (ISTI), National Research Council (CNR), Italy
Dimitris Rousidis	University of Alcalá, Spain
Rania Siatri	Alexander Technological Educational Institute (ATEI) of Thessaloniki, Greece
Miguel-Ángel Sicilia	University of Alcalá, Spain
Armando Stellato	University of Rome Tor Vergata, Italy
Sirje Virkus	Tallinn University, Estonia
Georgia Zafeiriou	University of Macedonia, Greece

Track on Metadata and Semantics for Open Repositories, Research Information Systems and Data Infrastructures

Special Track Chairs

Miguel-Ángel Sicilia	University of Alcalá, Spain
Armando Stellato	University of Rome Tor Vergata, Italy

Honorary Track Chairs

Imma Subirats	Food and Agriculture Organization (FAO) of the United Nations, Italy
Nikos Houssos	RedLink, Greece

Program Committee

Sophie Aubin	INRA (Institut National de la Recherche Agronomique), France
Thomas Baker	Sungkyunkwan University, Korea
Hugo Besemer	Wageningen UR Library, The Netherlands
Gordon Dunshire	University of Strathclyde, UK
Jan Dvorak	Charles University of Prague, Czech Republic
Jane Greenberg	Drexel University, USA
Siddeswara Guru	University of Queensland, Australia
Keith Jeffery	Keith G. Jeffery Consultants, UK
Nikolaos Konstantinou	University of Manchester, UK
Rebecca Koskela	University of New Mexico, USA
Jessica Lindholm	Malmö University, Sweden
Paolo Manghi	Institute of Information Science and Technologies (ISTI), National Research Council (CNR), Italy
Brian Matthews	Science and Technology Facilities Council, UK
Eva Mendez Rodriguez	Universidad Carlos III, Spain
Jochen Schirrwagen	Bielefeld University, Germany
Birgit Schmidt	University of Göttingen, Germany
Joachim Schöpfel	University of Lille, France
Kathleen Shearer	Confederation of Open Access Repositories (COAR), Germany
Chrisa Tsinaraki	European Commission, Joint Research Centre, Italy
Yannis Tzitzikas	University of Crete and ICS-FORTH, Greece

Zhong Wang Sun-Yat-Sen University, China
Marcia Zeng Kent State University, USA

Track on Metadata and Semantics for Digital Humanities and Digital Curation (DHC)

Special Track Chairs

Ernesto William De Luca	Georg Eckert-Institute – Leibniz-Institute for International Textbook Research, Germany
Paolo Bianchini	Università degli Studi di Torino, Italy

Program Committee

Wolf-Tilo Balke	TU Braunschweig, Germany
Elena Gonzalez-Blanco	Universidad Nacional de Educación a Distancia, Spain
Francesca Fallucchi	Guglielmo Marconi University, Italy
Ana Garcia-Serrano	ETSI Informatica – UNED, Spain
Steffen Hennicke	Georg Eckert-Institute – Leibniz-Institute for International Textbook Research, Germany
Ivo Keller	TH Brandenburg, Germany
Maret Keller	Georg Eckert-Institute – Leibniz-Institute for International Textbook Research, Germany
Andreas Lommatzsch	TU Berlin, Germany
Philipp Mayr	GESIS, Germany
Gabriela Ossenbach	UNED, Spain
Alessandra Pieroni	Guglielmo Marconi University, Italy
Christian Scheel	Georg Eckert-Institute – Leibniz-Institute for International Textbook Research, Germany
Lena-Luise Stahn	Georg Eckert-Institute – Leibniz-Institute for International Textbook Research, Germany
Armando Stellato	University of Rome Tor Vergata, Italy
Andrea Turbati	University of Rome Tor Vergata, Italy
Andreas Weiß	Georg Eckert-Institute – Leibniz-Institute for International Textbook Research, Germany

Track on Metadata and Semantics for Agriculture, Food and Environment (AgroSEM'17)

Special Track Chairs

Juliette Dibie	AgroParisTech and INRA, France
Liliana Ibanescu	AgroParisTech and INRA, France

Program Committee

Ioannis Athanasiadis	Wageningen University, The Netherlands
Patrice Buche	INRA (Institut National de Recherche Agronomique), France
Caterina Caracciolo	Food and Agriculture Organization (FAO) of the United Nations, Italy
Johannes Keizer	Food and Agriculture Organization (FAO) of the United Nations, Italy
Stasinos Konstantopoulos	NCSR "Demokritos", Greece
Claire Nédellec	INRA (Institut National de Recherche Agronomique), France
Ivo Jr. Pierozzi	Embrapa Agricultural Informatics, Brazil
Armando Stellato	University of Rome Tor Vergata, Italy
Maguelonne Teisseire	Irstea Montpellier, France
Jan Top	Wageningen Food & Biobased Research, The Netherlands
Robert Trypuz	John Paul II Catholic University of Lublin, Poland

Track on Metadata and Semantics for Knowledge IT Artifacts (KITA) in Professional Communities and Aggregations (KITA 2017)

Special Track Chairs

Fabio Sartori	University of Milano-Bicocca, Italy
Angela Locoro	University of Milano-Bicocca, Italy
Arlindo Flavio da Conceição	Federal University of São Paulo (UNIFESP), Brazil

Program Committee

Federico Cabitza	University of Milano-Bicocca, Italy
Luca Grazioli	ICteam SpA, Italy
Riccardo Melen	University of Milano-Bicocca, Italy
Aurelio Ravarini	Università Carlo Cattaneo – LIUC, Castellanza, Italy
Carla Simone	University of Siegen, Germany
Flávio Soares Corrêa da Silva	University of São Paulo, Brazil
Cecilia Zanni-Merk	Insa Rouen Normandie, France

Keynote Speaker

Dr. Trond Aalberg, Associate Professor

Department of Computer Science, Data and Artificial Intelligence Group - Norwegian University of Science and Technology

Dr. Trond Aalberg is Associate Professor at the Department of Computer Science, Data and Artificial Intelligence Group, Norwegian University of Science and Technology. His main research area is metadata and ontologies with a special emphasis on bibliographic information models, the transition of legacy data to new models and formats, and how these models can be utilized to improve search, exploration and reuse.

Keynote title: The path towards bibliographic ontologies and linked data

Libraries have a long tradition of exchange and reuse of data and were early adopters of common formats and standards for describing and coding bibliographic information. In the last decades, the community has faced the challenge of new models, formats and systems in order to adapt to the new digital environment and its requirements. The rather dramatic and hard to implement paradigm shift, that was introduced in the IFLA report on Functional Requirements for Bibliographic Records, was to introduce an entity-relationship perspective on the bibliographic domain and to model intellectual and artistic creations using entities reflecting different levels of abstraction.

This keynote will give an introduction to and contextualize the models that have been developed in the last decades, present challenges and results of the transformation and reinterpretation of legacy data, discuss quality issues and requirements that must be in place for this data to be of value when shared and reused as linked open data. Finally, the talk will present current research on how the models will enable new features for searching and exploration and thus increase the user experience and reuse value of the data.

Abstracts

General Session

Version Control and Change Validation for RDF Datasets

Manuel Fiorelli, Maria Teresa Pazienza, Armando Stellato, Andrea Turbati

Department of Enterprise Engineering,
University of Rome Tor Vergata
Via del Politecnico, 1, 00133 Rome, Italy
{fiorelli, pazienza, turbati}@info.uniroma2.it
stellato@uniroma2.it

The dynamic and distributed nature of the Semantic Web demands for methodologies and systems fostering collective participation to the evolution of datasets. In collaborative and iterative processes for dataset development, it is important to keep track of individual changes for provenance. Different scenarios may require mechanisms to foster consensus, resolve conflicts between competing changes, reversing or ignoring changes etc... In this paper, we perform a landscape analysis of version control for RDF datasets, emphasizing the importance of change reversion to support validation. Firstly, we discuss different representations of changes in RDF datasets and introduce higher-level perspectives on change. Secondly, we analyze diverse approaches to version control. We conclude by focusing on validation, characterizing it as a separate need from the mere preservation of different versions of a dataset.

Keywords: change management, version control, change validation, collaborative editing, RDF, knowledge bases, metadata.

Effect of Enriched Ontology Structures on RDF Embedding-Based Entity Linking

Emrah Inan and Oguz Dikenelli

Department of Computer Engineering,
Ege University 35100 Bornova, Izmir, Turkey

RDF embeddings are recently used in Entity Linking systems for disambiguation of candidate entities to match the best mention and entity pairs. In this study, we evaluate the effect of enriched ontology structures for disambiguation task when RDF embeddings are used to identify semantic relatedness between knowledge base concepts. We generate a domain-specific core ontology and put new components upon previous ontology structures. In this way, we obtain four different enriched structures and transform them into RDF embeddings. Then, we observe which enriched structure has more importance to enhance the overall performance of RDF embeddings-based Entity Linking approaches. We select two well-known knowledge-base-agnostic approaches, including AGDISTIS and DoSeR and adapt them into RDF embeddings-based entity disambiguation. Finally, a domain-specific

evaluation dataset is generated from Wikipedia to observe the effect of enriched structures on these adapted approaches.

Keywords: RDF embeddings, RDF2Vec, HITS, PageRank.

Cross-Querying LOD Datasets Using Complex Alignments: an application to agronomic taxa

Elodie Thiéblin[1], Fabien Amarger[1], Nathalie Hernandez[1], Catherine Roussey[2], and Cassia Trojahn Dos Santos[1]

[1] IRIT UMR 5505, 118 Route de Narbonne, F-31062 TOULOUSE CEDEX 9,
firstname.lastname@irit.fr
[2] Irstea, 9 avenue Blaise Pascal CS 20085 63178 Aubière,
firstname.lastname@irstea.fr

Farmers have new information needs to change their agricultural practices. The Linked Open Data is a considerable source of knowledge, separated into several heterogeneous and complementary datasets. This paper presents a process to query LOD datasets from a known ontology using complex alignments. The approach was applied on AgronomicTaxon, a taxonomic classification ontology, to query Agrovoc and DBpedia.

Keywords: query rewriting, complex alignments, agronomic sources, linked open data.

Deploying Metadata on Blockchain Technologies

Elena García-Barriocanal, Salvador Sánchez-Alonso, and Miguel-Ángel Sicilia

Computer Science Department, University of Alcalá,
Polytechnic building, Ctra. Barcelona km. 33.6
28871 Alcalá de Henares (Madrid), Spain
{msicilia, salvador.sanchez, elena.garciab}@uah.es

Metadata repositories and services provide support the key functions required by the curation of digital resources, including description, management and provenance. They typically use conventional databases owned and managed by different kinds of organizations that are trusted by their users. Blockchains have emerged as a means to deploy decentralized databases secured from tampering and revision, opening the doors for a new way of deploying that kind of digital archival systems. In this paper we review and evaluate the functions of metadata in that new light and propose an approach in which a blockchain combined with other related technologies can be arranged in a particular way to obtain a decentralized solution for metadata supporting key functions. We discuss how the approach overcomes some weaknesses of current digital archives, along with its important implications for the management and sustainability of digital archives.

Keywords: Blockchain, metadata, decentralization, provenance, trust, Ethereum, IPFS, BlockchainDB.

Creative Knowledge Environments Promotion through Case-Based Knowledge Artifacts

Fabio Sartori

Department of Computer Science, Systems and Communication (DISCo)
University of Milan - Bicocca
viale Sarca, 336
20126 - Milan (Italy)
sartori@disco.unimib.it

The adoption of case-based reasoning could be useful in the development of creative knowledge environments. In fact, it is one of the most suitable to reproduce decision making processes according to the reasoning by analogy paradigm. Given that analogies, and distant analogies in particular, are strictly connected to human creativity, case-based reasoning would result a good approach to provide knowledge artifacts with the capability to solve intrinsically creative problems. This is an important research topic in knowledge artifacts research: to this aim the paper will introduce CKS-Net, a conceptual and computational framework to manage with creative problems and domains, from both the theoretical and practical perspective.

Representation of Tensed Relations in OWL: a survey of philosophically-motivated patterns

Paweł Garbacz[1] and Robert Trypuz[2]

[1] Tadeusz Manteuffel Institute of History, Polish Academy of Sciences
[2] John Paul II Catholic University of Lublin, Faculty of Philosophy, Poland
{garbacz, trypuz}@kul.pl

The topic of this paper are the so-called tensed relations, i.e., those relations that hold between objects with respect to time. As tensed relations are not, almost by definition, binary relations, they need a special treatment in the case of such formal languages as OWL where only binary relations are explicitly expressible. We study in this paper a number of ways in which this expressivity constraint can be worked around focusing only on the solutions that seek their rationale in a philosophical argument of some sort. Besides Fleshing out the details of those patterns we compare them to one another to show their strengths and limitations in various usage scenarios.

A Data Exchange Tool Based on Ontology for Emergency Response Systems

Félix Simas[1,4], Rebeca Barros[1,2], Laís Salvador[1,2], Marian Weber[3], and Simone Amorim[4]

[1] Fraunhofer Project Center for Software and Systems Engineering at UFBA
[2] Federal University of Bahia (UFBA)
[3] Karlsruhe Institute of Technology (KIT)
[4] Federal Institute of Bahia (IFBA)
{felixneto, simone.amorim}@ifba.edu.br, rebecasbarros@dcc.ufba.br, laisns@ufba.br, marian.weber@student.kit.edu

Considering the potential of Emergency Response Systems (ERS) in combination with crowd-based information, this article presents a data integration solution within the scope of the RESCUER project: an ontology based data exchange solution to allow semantic interoperability between ERS of Command and Control Centers, referred to as Legacy Systems, and RESCUER. The solution is implemented by the Data Integration with Legacy Systems (DILS) component which can be used for the exchange of incident information by any ERS. As a result, we evaluated simulated emergency cases by sending RESCUER emergency reports to another ERS with the DILS.

Keywords: Emergency Response System, data Integration, data exchange, EDXL-RESCUER.

An Ontology Based Approach for Host Intrusion Detection Systems

Özgü Can, Murat Osman Ünalır, Emine Sezer, Okan Bursa, Batuhan Erdogdu

Ege University, Department of Computer Engineering, 35100 Bornova-Izmir, Turkey
ozgu.can@ege.edu.tr, murat.osman.unalir@ege.edu.tr, emine.sezer@ege.edu.tr, okan.bursa@ege.edu.tr, batuhanerdogdu@gmail.com

In recent years, cyber-attacks have emerged and these attacks result in serious consequences. In order to overcome these consequences, a fully-functioning and performance-improved intrusion detections systems are required. For this purpose, we used ontologies to provide semantic expressiveness and knowledge description for an intrusion detection system. In this work, a host intrusion detection system is implemented by using ontologies. The proposed system scans for malwares running on the operating system. Also, services and processes that are working on the system are scanned, and results are compared with the malware database. If any match occurs, the proposed system displays a malware list that matches with the information of that malware and where it is running. The proposed ontology based intrusion detection system aims to reduce the search time for malware scanning and to improve the performance of intrusion detection systems.

Keywords: Intrusion Detection System, Host Intrusion Detection, Ontology, Semantic Web.

An Approach for Systematic Definitions Construction Based on Ontological Analysis

Patricia Merlim Lima Scheidegger[1], Maria Luiza M. Campos[1], and Maria Cláudia Cavalcanti[2]

[1] Federal University of Rio de Janeiro, Rio de Janeiro, Brazil

patricia.merlim@ufrj.br; mluiza@ufrj.br

[2] Military Institute of Engineering, Rio de Janeiro, Brazil

yoko@ime.eb.br

This research motivation is to find a way to minimize the distance between business concepts and their respective representations in Information Technology artifacts, specially conceptual models. This distance leads to inconsistencies, ambiguities and implementation issues. Our approach is based on ontological analysis, which considers each concept according to its nature, capturing more precisely its essence and generally improving semantic richness and precision. Our main goal is to help in the process of systematic concepts definitions construction, contributing to generate more consistent definitions and associated conceptual modeling artifacts. The foundational ontology used as a theoretical reference is the Unified Foundational Ontology (UFO) which has been, in the last decade, successfully used to evaluate conceptual modeling languages and representations.

Keywords: conceptualization, definitions, conceptual modeling.

Track on Digital Libraries, Information Retrieval, Big, Linked, Social & Open Data

The Representation of Agents as Resources for the Purpose of Professional Regulation and Global Health Workforce Planning

Amy Opalek[1,2], Jane Greenberg[1]

[1] Metadata Research Center <MRC>, College of Computing & Informatics (CCI),
Drexel University, Philadelphia, PA, USA
{ao58, jg3243}@drexel.edu
[2] Foundation for Advancement of International Medical Education and Research (FAIMER),
Philadelphia, PA, USA

International migration of health professionals has been increasing in our globalized world, compounding a pressing need to improve information systems that confirm their qualifications and track health workforce volume. This paper reports on research to help address this need by introducing a framework for defining health professionals as agents. A case study and a categorical analysis of 11 metadata schemes was conducted. The results report and discuss three approaches to the representation of agents as either Access Point, Information Object, or Resource. The schemas that describe agents as resources best align with the representation of health professionals for the purpose of health workforce research.

Keywords: metadata, agent metadata, health workforce, HRIS for health.

Promoting Semantic Annotation of Research Data by their Creators: a use case with B2NOTE at the end of the RDM workflow

Yulia Karimova, João Aguiar Castro, João Rocha da Silva, Nelson Pereira, and Cristina Ribeiro

INESC TEC, Faculdade de Engenharia, Universidade do Porto,
Rua Dr. Roberto Frias, 4200-465 Porto, Portugal
{ylaleo, joaoaguiarcastro, joaorosilva, nelsonpereira1991}@gmail.com, mcr@fe.up.pt

Research data management is promoted at different levels with awareness actions carried out to encourage cooperation between researchers. However, data management requires tools to set the scene for researchers and institutions to disseminate the research data they produce. In this context good quality metadata play an important role by enabling data reuse. EUDAT is an European common data infrastructure, with integrated services for data preservation and dissemination. The TAIL project, at the University of Porto, proposes workflows based on Dendro, a collaborative environment that helps researchers prepare well described datasets and deposit them in a data repository. We propose a data deposit workflow use case for a small research project with emphasis in data annotation. Data is organized and described in Dendro; deposited in B2SHARE; and semantic annotation is performed with the new B2NOTE service from EUDAT.

Keywords: research data management, Dendro, B2NOTE, semantic annotation.

A Review of Practices for Transforming Library Legacy Records into Linked Open Data

Ya-Ning Chen

Department of Information and Library Science, Tamkang University,
New Taipei City 25137, Taiwan, R.O.C.
arthur9861@gmail.com

Current practices for transforming library legacy records into linked data for libraries were studied. The Linked Data Cookbook released by W3C was used as an analytical framework. A total of sixteen library linked data case studies focused on converting library catalogue data into linked open data were selected as subjects to analyze the details of transformation according to the following categories: identifying data, modeling data, naming with URIs, reusing existing terms, publishing human and machine readable descriptions, RDF con-version, license, download, host and announcement. It was found that although most tasks defined by the Linked Data Cookbook were adopted, some extensions and refinements were adopted to meet specific library-oriented requirements. Related issues including selection of data, selection of terms, 1-to-1 mapping principle and long-term preservation of library linked data are discussed.

Keywords: library legacy records, linked open data, library linked data.

Automatic Extraction of Correction Patterns from Expert-Revised Corpora

Giovanni Siragusa[1], Luigi Di Caro[1], and Marco Tosalli[2]

[1] University of Turin - Department of Computer Science Corso Svizzera 185, Turin, Italy
{siragusa,dicaro}@di.unito.it
[2] Nuance Communication Inc., Strada del Lionetto, 6, Turin, Italy
marco.tosalli@nuance.com

In this paper, we first present the task of automatically extracting correction patterns from texts which have been manually revised by domain experts. In real industrial scenarios, raw texts obtained via surveys or web crawling often require manual intervention to flatten word capitalization, punctuation, linguistic variability and entity naming. In this context, we propose a distributional and language-independent approach that learns revision rules that also manages errors introduced by the experts themselves. We extensively evaluated our approach on more than 300,000 expert-revised sentences.

Keywords: pattern extraction, natural language understanding, annotation learning, correction patterns.

Enabling Analysis of User Engagements Across Multiple Online Communication Channels

Zaenal Akbar, Anna Fensel, Dieter Fensel

Semantic Technology Institute (STI) Innsbruck, University of Innsbruck, Austria
{zaenal.akbar, anna.fensel, dieter.fensel}@sti2.at

The role of online communication channels, especially social media, has been developed from a platform for sharing information to a platform for influencing audiences. With the intention to reach the widest audience possible, organizations tend to distribute their marketing information to as many communication channels as possible. After that, they measure the performance of their marketing activities on every channel, where the typical measurement on how users perceived information is through engagement indicators. Measuring engagements across channels is challenging because the heterogeneity of engagement mechanism that can be performed by users on every channel. In this paper, we introduce a method to enable an analysis of those heterogeneous engagements which are distributed on multiple online communication channels. The solution consists of a conceptual model to uniformly representing user engagements on every channel. The model enables user engagements integration across channels, such that a more advanced user engagements analysis can be performed. We show how to apply our solution to analyze wide variety user engagements on popular social media channels from the tourism industry. This work brings us a step closer to realize an integrated multi-channel online communication solution.

Keywords: user engagement, multi-channel, data integration, data analysis.

Lost Identity – Metadata Presence in Online Bookstores

Tjaša Jug[1] and Maja Žumer[2]

[1,2] University of Ljubljana, Faculty of Arts, Department of Library and Information Science and Book Studies, Slovenia
Tjasa.jug@ff.uni-lj.si, Maja.zumer@ff.uni-lj.si

Book metadata plays an important role in discovering, identifying and selecting books in the online bookstores. While there is a link between good book description and sales, an insufficient description may make a book unfindable and therefore lost. There are many recommendations and guidelines regarding the book metadata to be included in the book description but not much is known about which information is actually available to end users. To get an overview of the current situation we conducted an expert study and examined American and Slovenian bookstores. We were mostly interested in the presence and the form of various metadata elements in online bookstores. The results show that all bookstores provide basic information, but many lack enhanced book information, which was evident especially in Slovenian sample. What is more, we found that metadata is mostly descriptive and does not allow users to navigate through the webpage and explore the collection. The results offer an insight in the actual situation on two book markets and open new research questions for publishers as well as for librarians.

Keywords: book metadata, online bookstores, expert study.

Collaborative Approach to Developing a Multilingual Ontology: a case study of wikidata

John Samuel

Université de Lyon, CPE Lyon, Lyon, France
john.samuel@cpe.fr

Last several years have seen the growing shift towards collaborative development of ontologies. Collaborative ontology development has become important particularly for large-scale projects involving multilingual contributors from different countries. Collaborators propose, discuss, create and modify ontologies and this whole process must be understood. In this article, Wikidata has been taken as an example to understand how community-driven approach is used to develop a multilingual ontology and in the subsequent building of a knowledge base.

New Generation Metadata Vocabulary for Ontology Description and Publication

Biswanath Dutta[1], Anne Toulet[2], Vincent Emonet[2] and Clement Jonquet[2,3]

[1] Documentation Research and Training Centre (DRTC)
Indian Statistical Institute, Bangalore, India
bisu@drtc.isibang.ac.in
[2] Laboratory of Informatics, Robotics and Microelectronics of Montpellier (LIRMM)
CNRS & University of Montpellier, France
{anne.toulet, emonet, jonquet}@lirmm.fr
[3]Center for Biomedical Informatics Research (BMIR)
Stanford University School of Medicine, USA

Scientific communities are using an increasing number of ontologies and vocabularies. Currently, the problem lies in the difficulty to find and select them for a specific knowledge engineering task. Thus, there is a real need to precisely describe these ontologies with adapted metadata, but none of the existing metadata vocabularies can completely meet this need if taken independently. In this paper, we present a new version of *Metadata vocabulary for Ontology Description and publication*, referred as MOD 1.2 which succeeds previous work published in 2015. It has been designed by reviewing in total 23 standard existing metadata vocabularies (e.g., Dublin Core, OMV, DCAT, VoID) and selecting relevant properties for describing ontologies. Then, we studied metadata usage analytics within ontologies and ontology repositories. MOD 1.2 proposes in total 88 properties to serve both as (i) a vocabulary to be used by ontology developers to annotate and describe their ontologies, or (ii) an explicit OWL vocabulary to be used by ontology libraries to offer semantic descriptions of ontologies as linked data. The experimental results show that MOD 1.2 supports a new set of queries for

ontology libraries. Because MOD is still in early stage, we also pitch the plan for a collaborative design and adoption of future versions within an international working group.

Keywords: metadata vocabulary, ontology metadata, semantic description, ontology repository, ontology reuse, ontology selection, ontology relation.

Track on Cultural Collections and Applications

Enriching Media Collections for Event-Based Exploration

Victor de Boer[1,4], Liliana Melgar[2,4], Oana Inel[1], Carlos Martinez Ortiz[3], Lora Aroyo[1], and Johan Oomen[4]

[1] Department of Computer Science, Vrije Universiteit Amsterdam, Amsterdam, the Netherlands
{v.de.boer, oana.inel, lora.aroyo}@vu.nl
[2] Universiteit van Amsterdam, Amsterdam, the Netherlands
melgar@uva.nl
[3] eScience Center, Amsterdam, the Netherlands
c.martinez@esciencecenter.nl
[4] Netherlands Institute for Sound and Vision, Hilversum, the Netherlands
joomen@beeldengeluid.nl

Scholars currently have access to large heterogeneous media collections on the Web, which they use as sources for their research. Exploration of such collections is an important part in their research, where scholars make sense of these heterogeneous datasets. Knowledge graphs which relate media objects, people and places with historical events can provide a valuable structure for more meaningful and serendipitous browsing. Based on extensive requirements analysis done with historians and media scholars, we present a methodology to publish, represent, enrich, and link heritage collections so that they can be explored by domain expert users. We present four methods to derive events from media object descriptions. We also present a case study where four datasets with mixed media types are made accessible to scholars and describe the building blocks for event-based proto-narratives in the knowledge graph.

A Semantic Web Case Study: representing the Ephesus Museum collection using Erlangen CRM Ontology

Tuğba Özacar, Övünç Öztürk, Lobaba Salloutah, Fulya Yüksel, Baraa Abdülbaki, and Elif Bilici

Department of Computer Engineering, Manisa Celal Bayar University, 45140, Manisa, Turkey
tugba.ozacar@cbu.edu.tr, ovunc.ozturk@cbu.edu.tr, lobaba.salloutah@ogr.cbu.edu.tr, fulya.yuksel@ogr.cbu.edu.tr, baraa.abdulbaki@ogr.cbu.edu.tr, elif.bilici@ogr.cbu.edu.tr

Cultural heritage has recently become an important application area for Semantic Web technologies. Semantic Web technologies and ontologies provide a solution for intelligent integration of heterogeneous data about the cultural heritage. The objective of this paper is the construction of an ontology for the cultural heritage related to Selcuk region in Western Turkey. We use a subset of the Erlangen CRM as our ontology schema, then we populate the ontology with 814 objects in the Ephesus Museum. One of the objectives of this work is to integrate the ontology with other projects which use Erlangen CRM as ontology schema. Therefore, we present an integration case study that aggregates content from Ephesus Museum and British Museum.

The Semantic Enrichment Strategy for Types, Chronologies and Historical Periods in searchculture.gr

Haris Georgiadis, Agathi Papanoti, Maria Paschou, Alexandra Roubani, Despina Hardouveli, Evi Sachini

National Documentation Centre / National Hellenic Research Foundation Athens, Greece
{hgeorgiadis, apapano, mpasxo, arouba, dxardo, esachin}@ekt.gr

Most aggregators face challenges regarding searchability, discoverability and visual presentation of their content due to metadata heterogeneity. We developed an innovative metadata enrichment and homogenization scheme that is both effective and user-friendly and we embedded it in the ingestion workflow of searchculture.gr, the cultural heritage aggregator of National Documentation Centre (EKT). Two key components of the enrichment scheme are semantics.gr, a platform for publishing vocabularies that contains a tool for massive semantic enrichment, and a parametric tool embedded in the aggregator for chronological normalization. We enriched and homogenized the aggregated content with respect to types and chronological information which subsequently allowed us to develop advanced multilingual search and browsing features, including hierarchical navigation on types and historical periods, searching and faceting on type, time span and historical period, a tag cloud of types and an interactive timeline.

Keywords: aggregator, semantic enrichment, linked data, automatic categorization, vocabularies, thesauri, cultural heritage, historical periods, time-driven search, temporal coverage, timeline.

Linked Data in Libraries' Technical Services Workflows

Philip E. Schreur and Nancy Lorimer
Stanford University, Stanford CA 94305, USA

Linked Data for Production (LD4P) is a collaborative project between six institutions (Columbia, Cornell, Harvard, the Library of Congress, Princeton, and Stanford) to begin the transition of the production workflows of their libraries Technical Services Departments to ones rooted in Linked Open Data (LOD). Each institution is focused on a different domain or facet of the problem to move us together as a group more quickly. As a whole, the six institutions will focus on four main areas of development. First will be the establishment of the ability to create linked open data communally. Second, in collaboration with external standards organizations such as the Program for Cooperative Cataloging and linked data projects such as BIBFLOW, will be the establishment of common procedures and protocols for the creation of library metadata as linked data. Third will be the expansion of the BIBFRAME ontology to better encompass subject domains such as art and music. And last will be the transition of a selection of current library workflows to ones based in linked open data. The projects will make use of a collection of preliminary tools and adopt them for production work in their individual environments and, through feedback, assist in the development of the tools.

Keywords: BIBFRAME, linked Data, ontology, cataloging.

The Combined Use of EAD and METS for Archival Material : an integrated toolkit

Ricardo Eito-Brun

Universidad Carlos III de Madrid, c/ Madrid 126, Getafe, Madrid, Spain
reito@bib.uc3m.es

Information professionals have at their disposal a complex ecosystem of standards and tools designed with specific focus by different user groups. In the case of the Cultural Heritage projects, there are different metadata schemas aimed to support the efficient creation, storage and distribution of content and digital assets. But the use of these metadata schemas, and the software tools that make their use and deployment possible, are sometimes restricted to well-limited areas. As an example, the EAD schema is widely used by archivists for describing fonds and records, TEI is mainly applied for encoding textual corpus, and METS is bounded to the digitization of ancient books. A major permeability and reuse of metadata schemas and tools in different contexts is needed, as information professionals and user communities can leverage their capability to exchange and disseminate assets, gain independence from proprietary software solutions and platforms and make possible an unlimited, global access to our communities' Cultural Heritage. This paper describes the development of a technical solution that integrates the use of EAD with METS for the digitization and description of archival records. With the proposed solution archivists obtain the benefits of both standards using a common, integrated process and tools.

Keywords: records management, EAD, Mets, metadata integration.

Track on European and National Projects

Developing Quiz Games Linked to Networks of Semantic Connections among Cultural Venues

Abdullah Daif[1], Ahmed Dahroug[2], Martín López-Nores[1], Alberto Gil-Solla[1], Manuel Ramos-Cabrer[1], José Juan Pazos-Arias[1], and Yolanda Blanco-Fernández[1]

[1] AtlantTIC Research Center, Department of Telematics Engineering, University of Vigo, Spain
adaif@uvigo.es, {mlnores, agil, mramos, jose, yolanda}@det.uvigo.es
[2] Arab Academy for Science, Technology and Maritime Transport, Egypt
adahroug_87@aast.edu

Existing general-purpose and domain-specific resources of the Semantic Web provide the foundations to discover connections between any two concepts of interest. The EU project CROSSCULT seeks to exploit that possibility in order to spur a change in the way citizens appraise history and culture, by means of web and mobile apps that will let them explore cross-border interconnections among cultural venues and their collections of heritage items. In this paper, we present a tool for experts to collaboratively develop the networks of semantic associations that will drive the gameplay or the storytelling of the apps. This tool includes reasoning aids to discover associations, to identify the most relevant ones in relation to selected topics, and to develop quiz tests involving the chosen entities (heritage items, characters, events or locations).

Keywords: cultural heritage, semantic associations, association discovery and ranking, reflective topics.

Metadata for Nanotechnology: interoperability aspects

Vasily Bunakov[1] and Brian Matthews[1]

[1] Science and Technology Facilities Council, Harwell OX11 0QX, UK
{vasily.bunakov, brian.matthews}@stfc.ac.uk

The work outlines the landscape of emerging metadata models for nanotechnology. A gap analysis and possible cross-walks for a few metadata recommendations are presented. The role of interoperability in the design of metadata for nanotechnology is discussed.

Keywords: nanotechnology, metadata models, metadata interoperability.

Data, Metadata, Narrative. Barriers to the Reuse of Cultural Sources

Jennifer Edmond (0000-0001-9991-1637) and Georgina Nugent Folan (0000-0002-6216-9317)

Trinity College Dublin, Dublin, Ireland

edmond@tcd.ie & nugentfg@tcd.ie

The networking of objects facilitated by the Internet of Things isn't new. Every object that is catalogued for display within a GLAM institution is assigned entry-level data, along with further data layers on that object that each interactive agent (researcher) will draw upon to create their research narratives, irrespective of their disciplinary background or bias. Within the community of researchers working with cultural data in particular, the desire to compare and aggregate diverse sources held together by a thin red thread of potential narrative cohesion, is only increasing. This poses challenges to information retrieval and contextualization in the digital age, it forces us to reassess the value and cost of metadata, and the consequences that accompany the use and reuse of digital data in a humanities or cultural research context. This paper discusses a number of the key barriers to the digital representation of complex cultural data and presents the preliminary findings and recommendations of the EU Commission's Horizon 2020 funded KPLEX project (kplexproject.eu) in the field of knowledge complexity and cultural data.

Keywords: narrative, data, metadata, cultural computing, digital humanities.

Validating the Ontology-Driven Reference Model for the Vocational ICT Curriculum Development

Mohammad Hadi Hedayati[1] and Mart Laanpere[2]

[1] Computer Science Faculty, Kabul University, Afganistan

hdhedayati@gmail.com

[2] School of Digital Technologies, Tallinn University, Estonia

martl@tlu.ee

In the globally standardized domain of ICT, the vocational education curricula have to balance the requirements of fast-evolving international standards with unique local cultural traditions and socio-economic needs. This paper summarizes the results of a case study that tackled this challenge by developing and validating a reference model for ontology-driven curriculum development in the context of vocational ICT education in developing countries a case study about Afghanistan. We have demonstrated that even in the most challenging socio-economic and cultural context, semantic technologies have potential to represent, organize, formalize and standardize the knowledge in ICT domain so that it can be built, shared and reused in the curriculum process by different stakeholders like teachers, employers, alumni and deans. The final stage of our iterative design-based research focuses on development and validation of Ontology-Driven Curriculum (ODC) reference model based on results of our previous research.

Keywords: ICT, vocational education, curriculum, ontology, semantic web, metadata, knowledge management, knowledge sharing, job market, reference model.

Track on Open Repositories, Research Information Systems and Data Infrastructures

Building Scalable Digital Library Ingestion Pipelines Using Microservices

Matteo Cancellieri (0000-0002-9558-9772), Nancy Pontika (000-0002-2091-0402), Samuel Pearce (0000-0001-5616-7000), Lucas Anastasiou (0000-0002-1587-5104), and Petr Knoth (0000-0003-1161-7359)

CORE, The Open University, Milton Kenyes, MK7 6AA, UK,
theteam@core.ac.uk
WWW home page: https://core.ac.uk/

CORE, a harvesting service offering access to millions of open access research papers from around the world, has shifted its harvesting process from following a monolithic approach to the adoption of a microservices infrastructure. In this paper, we explain how we rearranged and re-scheduled our old ingestion pipeline, present CORE's move to managing microservices and outline the tools we use in a new and optimised ingestion system. In addition, we discuss the inefficiencies of our old harvesting process, the advantages, and challenges of our new ingestion system and our future plans. We conclude that via the adoption of microservices architecture we managed to achieve a scalable and distributed system that would assist with CORE's future performance and evolution.

Keywords: harvesting, repositories, microservices, infrastructure, software, architecture.

Semantic Attributes for Citation Relationships: creation and visualization

Sergey Parinov (orcid.org/0000-0001-8333-2657)

Central Economics and Mathematics Institute of RAS, Moscow, Russia
Russian Presidential Academy of National Economy and Public Administration, Moscow
sparinov@gmail.com

This paper presents a method to process a content of research papers in binary PDF format at a server side that gives research information systems new features of citation content analysis. This method efficiently generates JSON versions of PDF documents that allows an easier recognition of papers' references, in-text references, citation context, etc. As a result, one can parse an extended set of citation data, including a location of citations in a research paper's structure, frequency of mentioning for the same references, style of reference mentioning and so on. Based on these data we upgrade traditional citation relationships by adding some semantic attributes. Formatting these semantic data according W3C Web Annotation Data Model and integrating the data with some annotation tools, we visualize citation relationships, its semantic attributes and related statistics as annotations for readers of PDF documents from a research information system.

Keywords: research information system, PDF.js, PDF to JSON conversion, citation relationships, semantic attributes, citation content analysis, visualization.

Toward a Metadata Framework for Sharing Sensitive and Closed Data: an analysis of data sharing agreement attributes

Sam Grabus[1], and Jane Greenberg[2]

[1,2] Metadata Research Center <MRC>, College of Computing & Informatics (CCI), Drexel University, Philadelphia, PA, USA

{sam.grabus, janeg}@drexel.edu

Legal and policy-oriented restrictions often hamper if not inhibit well-intended efforts to share sensitive or restricted data. The research reported on in this paper is a part of a larger initiative to develop a prototype system for automatically generating data sharing agreements that address privacy, legal concerns, and other restrictions. A content analysis was conducted, examining a sample of 26 data sharing agreements. The results include 6 high level categories, 15 mid-level attributes, and over 90 lower-level specific attributes, a portion of which can help to expeditiously support the automatic development of data sharing agreements. The paper presents background information, research questions and methods, results, and a discussion. The conclusion summarizes our results and identifies next steps.

Keywords: metadata, closed data, data sharing agreements, restricted data, privacy and data sharing.

Track on Digital Humanities and Digital Curation (DHC)

Battle without FAIR and Easy Data in Digital Humanities: an empirical research on the challenges of open data and APIs for the James Cook dynamic journal

Go Sugimoto[1]

[1] Austrian Centre for Digital Humanities, Austrian Academy of Sciences, Vienna, Austria
Go.Sugimoto@oeaw.ac.at

There are theoretical and technical challenges for Digital Humanities scholars to develop and use Application Programming Interfaces (APIs). Whilst data owning culture in our institutional settings seems to hinder truly interdisciplinary research, the emergence of Linked Open Data implies the increasing opportunities of distributed-data research. This article is based on an API application to address those issues in the context of Open Data. James Cook Dynamic Journal was created to assist users to study the Cook's journal, integrating various sets of APIs which facilitate full-text search, Named Entity Recognition, and map views. The development revealed some critical issues of data federation and processing automation. The standardization of data structure and the development of user-friendly GUI tools would significantly increase the value of APIs. Taking recent initiatives into account, the paper also proposes "Easy Data" to liberate Open Data for a wider spectrum of users outside the programmer community.

Keywords: API, digital humanities, James Cook, open data, FAIR principles, Easy Data.

Enrichment of Accessible LD and Visualization for Humanities: MPOC model and prototype

Alicia Lara-Clares, Ana Garcia-Serrano and Covadonga Rodrigo

ETSI Informática
Universidad Nacional de Educación a Distancia – UNED Madrid, Spain
{alara, agarcia, covadonga}@lsi.uned.es

In this paper, it is presented how to enrich available linked data to facilitate the work of experts in the Musacces project (http://www.musacces.es/). First step has been implemented in the current version of a prototype MPOC1, that allows the access and visualization of information related with the Spanish Prado Museum (the target of the project). Main aim of the project is to work with semantic technologies in two ways: the management of available data from different resources creating in a first step a "local repository" for research purposes. The implemented functionalities are based on a crawler that extracts all the available metadata from an artwork and its artist, and on an enrichment component allowing the collaboration of the experts, in order to access and visualize the information guided by the user.

The second step is going to be the use of a semantic graph that can assist the experts (humanists) not only to create a textual-narrative that explains distinct stages of history but to create a management

model that allows interaction with the semantic objects (and its relations) that make up the narration and its presentation modes (textual, sign language, etc.). This way, personalized narratives for certain groups of society can be created by the experts and being visualized from different points of view (storylines, maps, events lists…) as well as different presentation to the public, taking their abilities into consideration.

Keywords: museum, digital curation, linked data, narratives, semantic graph.

Posters

JabotG: Extending the Herbarium Dataset Frontiers

Felipe Alves de Oliveira[1,2], Yasmmin Cortes Martins[3], Diogo S. B. Rocha[2], Marinez Ferreira de Siqueira[2], Luís Alexandre Estevão da Silva[2], Raquel L. Costa[3], Ronaldo Ribeiro Goldschmidt[1], and Maria Cláudia Cavalcanti[1]

[1] Instituto Militar de Engenharia (IME), Brazil
{yoko; ronaldo.rgold}@ime.eb.br
[2] Instituto de Pesquisas Jardim Botânico do Rio de Janeiro (IPJBRJ), Brazil
{felipealves; marinez; estevao}@jbrj.gov.br, diogosbr@gmail.com
[3] Laboratório Nacional de Computação Científica (LNCC), Brazil
{yasmmincortes; quelopes}@lncc.br

Abstract: Herbaria around the world have been collecting and curating data throughout the years. Many of them maintain Web portals to provide online data from herbarium specimens and related collections, in addition to their images. More recently, Open Governmental Data Portals have been the standard solution for government institutions willing to put their public data available for the society. However, publishing data, which originally is stored in relational databases, into open linked data initiatives is not an easy task. This work reports on the efforts and the benefits of publishing one of the data collections of the Rio de Janeiro Botanical Garden Research Institute into an RDF graph, named JabotG. A set of queries illustrates the analytical potential of the new format, and how two related datasets can be queried to provide new interesting and useful insights.

Keywords: web of data, open linked data, RDF graph, biodiversity.

1 Introduction

Brazil is a country in which there is one of the richest biodiversities in the world, with about 46,430 different plant species recognized by Brazil's Flora [1]. This natural treasure is constantly endangered by human actions, such as habitat degradation, deforestation, pollution, climate change and biological invasions. It is a matter of national defense to the country to keep track and catalog such biodiversity, for their relevance to the environment in maintaining ecosystem services and to preserve the water resources as well as the food and health industry. In this context the herbaria play a paramount role, maintaining registries of botanical collections. The Rio de Janeiro Botanical Garden Research Institute (IPJBRJ) is responsible for maintaining exsiccates (dehydrated specimens) of botanical species. In addition, IPJBRJ also maintains the Jabot database [11] that includes the digitalization of this and other IPJBRJ collections, as well as digitalized data of many other Brazilian herbaria. Part of the data is available for public access through a Web site[4], for ad-hoc consultation. Many herbaria around the world provide similar services [2,4]. Through these databases, it is possible to extract useful information to provide a valuable resource for ecologists, evolutionary biologists, and conservationists. Further improvements associated with this data include new methods to objectively plan future survey work of endangered species; incorporation of collectors' field notes; and methodological research like ecological niche modeling, data mining, identification of species co-occurrence, among others [9].

Herbaria' databases are usually focused on a specific set of researchers, botanic areas, and occurrences. The Global Biodiversity Information Facility (GBIF) [12] is an important world funded initiative that addresses the integration of almost a thousand databases, covering 1,720,142 species.

It provides an open-data research infrastructure to access data about these species, reinforcing the use of standard vocabularies such as the Darwin Core[5].

More recently, Open Governmental Data Portals (ODPs) have emerged as the standard solution for governments willing to put their public data available to the society [13]. In Brazil, the SiBBr6 is the ODP that focus on data about the Brazilian biodiversity. They work in association with GBIF and use their specific standards to publish data. Jabot is one of the databases that publishes data on GBIF.

Another important initiative in the direction of data integration is the Web of Data [7]. The idea is to build a global data space containing billions of described data, through which it is possible to navigate. In this proposal, data items/resources are represented as nodes of a graph, using the Resource Description Framework[7] (RDF), a W3C Recommendation. According to it, each resource should be identified by accessible URIs (Uniform Resource Identifiers). A resource may be connected to another resource, forming a triple (node-edge-node), and connected triples form the graph.

To enable navigation through datasets, they must include data resources that point to other datasets. The Linking Open Data[8] (LOD) initiative is one of the main efforts that has contributed to the growth of the Web of Data. It provides a set of best practices that should be adopted by data publishers to facilitate the linking of data. One of these practices is to use standard or controlled vocabularies to form RDF assertions.

To the best of our knowledge, herbaria around the world did not report on publishing their data in the Web of Data. This work main contribution is the description of the choices that we made on the transformation of Jabot relational data into an RDF graph, such as the use of the Darwin Core vocabulary for modeling data. In addition, it shows the benefits of RDF format with respect to data integration, through some useful queries that combine data from multiple datasets.

This work is organized as follows. Next section presents some related initiatives on vocabularies and ontologies. Section 3 describes the modeling and building of JabotG dataset, reusing related vocabularies. Section 4 describes the IBGE/INDE dataset, and presents some queries over the integrated datasets. Then, we conclude the paper with some final discussion and future work.

2 Related Vocabularies, Taxonomies and Ontologies

As mentioned before, one of the best practices to publish data on the Web of Data is to use uniform vocabularies. There are some initiatives related to herbarium data that are already available to facilitate their publishing process and are described as follows.

The Darwin Core (DwC) vocabulary is the most used standard for biodiversity data interchange [14]. It was originally conceived to describe biological specimens, and their occurrence in time/space, as well as the evidence of such events in (physical/digital) collections. The main idea was to provide a stable reference standard that could be easily reused in many contexts, such as for integrating data from herbaria and museums. As mentioned before, Jabot and many other herbarium databases have their data published according to DwC, into the GBIF portal.

[4] http://jabot.jbrj.gov.br/
[5] http://rs.tdwg.org/dwc/
[6] http://www.sibbr.gov.br/
[7] https://www.w3.org/RDF/
[8] http://lod-cloud.net/

DwC is a collection of terms with a clear semantics that can be easily understood by people and machines. It allows researchers to describe biodiversity mainly by reporting the rate at which

specimens occur in nature. Documents for guiding its usage are available such as Baskauf (2016) [6][5]. They describe how to use the vocabulary, not only by defining the meaning of each term but also by defining them according to the RDF language elements. Triples of biodiversity information can be generated in RDF using the DwC vocabulary.

The example of Figure 1 was extracted from the DwC guide. It uses the DwC vocabulary and RDF syntax to describe a specimen identification, which was identified by Takuma Yun, and which is classified as a yaeyamaensis of the genus Hersilia and family Hersiliidae. All tags/terms preceded by dwc belong to the DwC vocabulary. According to the RDF syntax, a triple is composed of a subject, a predicate and an object. In this example, the Description tag expresses five triples. It takes the specimen (http:...503A4527D009) as the subject, and for each predicate, it associates an object. The predicate dwc:identifiedBy is associated to the literal Takuma Yun, forming one of the triples.

There are initiatives that propose a taxonomy (hierarchy of terms used to classify specimens), such as The Plant List [3] (TPL).TPL is a list of all known plant species, which comprises vascular plants and Bryophytes. It was the result of the collaboration of the Royal Botanic Gardens Kew [2] and the Missouri Botanical Garden [4]. This list has more than a million scientific names of species, more than 17,000 genus, and more than 600 families [3]. It also covers all known synonyms for these species. To the best of our knowledge, they did not provide RDF or OWL format downloading of this list.

```
<rdf:Description rdf:about="http://museum.or.jp/...503A4527D009">
        <rdf:type rdf:resource="http://rs.tdwg.org/dwc/terms/Identification"/>
        <dwc:identifiedBy>Takuma Yun</dwc:identifiedBy>
        <dwc:family>Hersiliidae</dwc:family>
        <dwc:genus>Hersilia</dwc:genus>
        <dwc:specificEpithet>yaeyamaensis</dwc:specificEpithet>
</rdf:Description>
```

Fig. 1. Example of DwC usage in RDF syntax

A similar taxonomy, which is already in use by Jabot database, is the Brazilian Flora 2020[9]. It comprises the list of Plants and Fungi of Brazil. It is possible that some species in the TPL may be the same as some species in the Brazilian Flora list, although labeled differently. To date, the Brazilian Flora catalog contains 46,430 species: 4751 Algae, 33053 Angiosperms, 1553 Bryophytes, 5722 Fungi, 30 Gymnosperms and 1321 Ferns and Lycophytes.

Another interesting initiative is the Planteome Project, which is a centralized online informatics portal and database, consisting of a suite of reference ontologies for plants, an associated corpus of plant genomics and phenomics data, and tools for data analysis and annotation [8]. The Planteome ontologies are the Plant Ontology[10] [10] (PO), the Plant Trait Ontology (TO), the Plant Environment Ontology (EO) and the Plant Stress Ontology (PSO). These ontologies focus on specific plant characteristics, such as plant anatomy or development stages.

[9] http://floradobrasil.jbrj.gov.br/
[10] http://www.plantontology.org/

3 Modeling and Building JabotG Dataset

The Jabot database was developed using PostgreSQL DBMS, and it comprises more than 122 tables, 42 views, 19 triggers and 95 functions. It was designed to store data from 13 scientific collections, like exsiccates (vouchers), woods, seeds, DNA, among others. However, to generate a first version of the graph, we chose to work with only one collection, the exsiccates. The complete database has a total of 765,000 tuples, while the exsiccates collection has about 690,000 tuples. In addition, to reduce complexity, just a subset of tables and attributes of the Jabot schema was used: tables related to specimens occurrences at specific locations, and their association to the corresponding species.

```
<rdf:RDF xmlns:dwc="http://rs.tdwg.org/dwc/terms/"
xmlns:foaf="http://xmlns.com/foaf/0.1/"
xmlns:jbg="http://jabot.jbrj.gov.br/jabotg/terms/"
xmlns:rdfs="https://www.w3.org/TR/rdf-schema/">
<rdf:Description rdf:about="jbg:RB00809155">
   <rdf:type rdf:resource="jbg:Occurrence"/>
   <dwc:recordedBy rdf:resource="jbg:RCForzza"/>
   <dwc:associatedTaxa>
      <rdf:Description rdf:about="jbg:alba">
         <rdf:type rdf:resource="jbg:SpecificEpithet"/>
         <dwc:genus>
            <rdf:Description rdf:about="jbg:Aechmea">
               <rdf:type rdf:resource="jbg:Genus"/>
               <dwc:family
          rdf:resource="jbg:BROMELIACEAE"/></rdf:Description>
          </dwc:genus></rdf:Description>
    </dwc:associatedTaxa></rdf:Description> ...
```

Fig. 2. JabotG Triples

The generated graph, extracted from Jabot data, has a total of 523,537 nodes and 2,547,952 edges, with attributes and properties. Figure 4 shows the graph schema (Dt JabotG) that was developed to represent the data. According to this schema, each specific epithet, each genus and each family are now represented as nodes in the graph. Also, each occurrence of a specimen, as well as its location and the researcher who recorded it, are represented as a node in the graph. The schema directed edges indicate that instances of each of these node types may be connected to other node type instances, through typed directed edges. For instance, associations between a specimen occurrence and its taxon, were represented using the dwc:associatedTaxa edge in the graph.

Each directed edge forms an RDF triple, which is composed by a source node and a target node. In terms of the RDF elements, the source is the subject, the target is the object, and the edge is the predicate, that applied to the subject leads to the object, forming an RDF triple. In the example of Figure 2, the RB00809155 occurrence is a specimen recorded by RCForzza, and it is associated to the alba specific epithet, forming two RDF triples.

Note that the DwC vocabulary was used to represent many of the resources (nodes) of the graph. The idea was to reuse as much as possible of this vocabulary, since it is recommended by data integration initiatives such as GBIF. In addition, we also used the FOAF[11] and the RDFS[12] vocabularies, and some classes and properties of our own, which are prefixed with jbg namespace.

We used the dwc:associateTaxa property to connect a specimen found at a given place, at a given time (dwc:occurrence), to a specific taxon (dwc:Taxon). In our work, we aim to represent each taxon as a resource in order to enable their association to other taxonomies. Thus, we need to represent them explicitly, at their specific taxonomic levels (e.g. genus, family or specific epithet). Therefore, specific node types (jbg:Family, jbg:Genus, jbg:SpecifiEpithet) were created as subclasses (rdfs:subClassOf) of dwc:Taxon. These node types are connected by dwc:family and dwc:genus properties. The Brazilian Flora taxonomy is used to instantiate the dwc:Taxon subclasses.

Terms of the FOAF social vocabulary were used to represent the researchers or people involved in expeditions and collection of specimens (foaf:Person, foaf:interest). These researchers usually take part on expeditions (jbg:Expedition) to visit (jbg:visit) some location(dwc:location), that in the case of Jabot, must be specialized in the following classes: geopolitical or conservation unit (jbg:GeopoliticalUnit, jbg:ConservationUnit).

In order to explore and visualize useful queries, JABOT data was also loaded into a graph database system called Neo4J[13]. Queries in the JabotG dataset are quite intuitive. Some examples of queries are presented as follows using the Neo4J query language (CYPHER), and its corresponding results. The first one explores the idea of social analysis. It retrieves researchers that have similar research interests, and thus that could, potentially, collaborate with each other. In other words, it tries to suggest social links between researchers Figure 3.

Note that such queries could be expressed in SQL and submitted to the Jabot database. However, expressing them in SQL are quite more complex. Besides, such queries may involve many joins, which may lead to performance issues.

4 IBGE/INDE Dataset

In parallel to the modeling and building of JabotG, another dataset with complementary data has been under preparation. It correlates locations with their corresponding biome, climate, vegetation, and soil. This dataset was built using data generously provided by the Brazilian Institute of Geography and Statistics[14] (IBGE) and the National Infrastructure for Spacial Data[15] (INDE). The idea of creating the IBGE/INDE dataset was to demonstrate the usefulness of integrating datasets, as graphs in the Web of Data. Figure 4 shows that JabotG references IBGE/INDE dataset, through its location nodes, i.e., each JabotG location node points to a similar node in the other dataset.

In order to illustrate how useful it is that both datasets are in the Web of Data, naturally integrated, some queries that cross their frontiers are presented as follows. The first one, shown in Figure 5, presents what are the different climates in which a given family (BROMELIACEAE) occurs. The second one (Figure 6) presents an aggregation that indicates the frequency with which each family occurs in each biome, such as Atlantic Forest (Mata Atlântica) and Brazilian Savanna (Cerrado).

[11] http://xmlns.com/foaf/spec/
[12] https://www.w3.org/TR/rdf-schema/
[13] https://neo4j.com/
[14] http://www.ibge.gov.br/
[15] http://www.inde.gov.br/

```
MATCH (n:Person{recordBy:'R.C.Forzza'})-[:RecordedBy]-(o)-[:associatedTaxa]
      ->(s)<-[:associatedTaxa]-()-[:RecordedBy]-(ParCollector)
```

RETURN DISTINCT ParCollector.recordBy **ORDER BY** ParCollector.recordBy

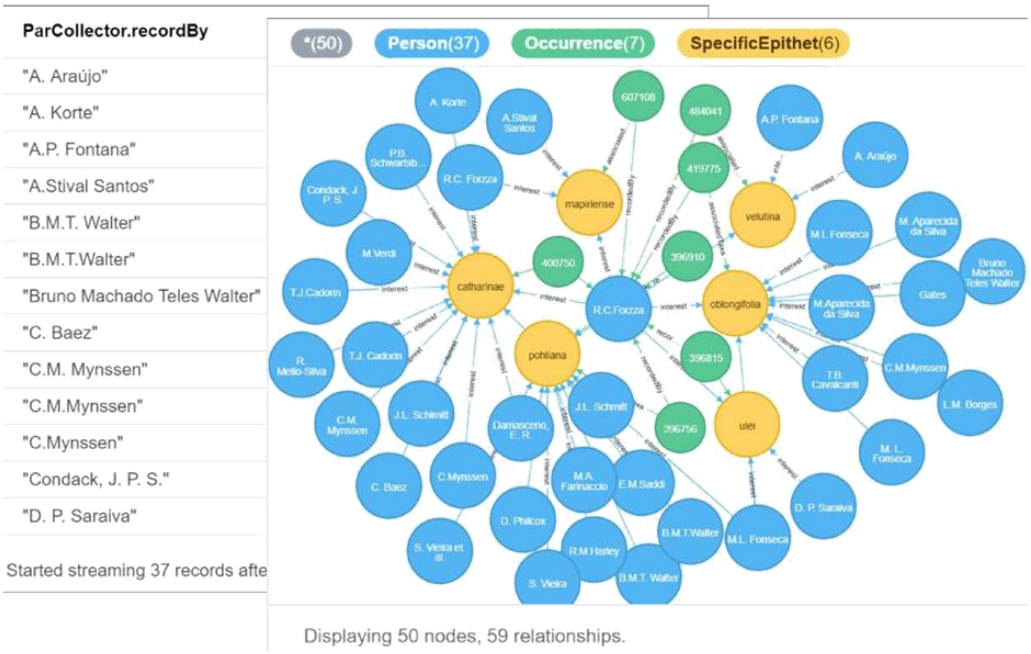


Fig. 3. Identifying researchers that share family interests with Rafaela Forzza

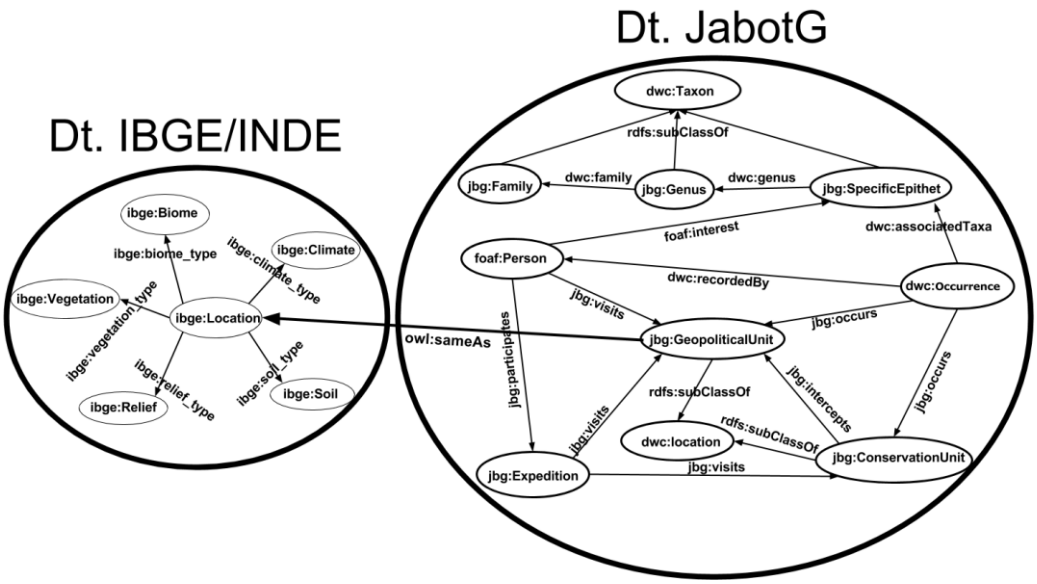


Fig. 4. Integrated Datasets

MATCH (f:Family{family: 'BROMELIACEAE'})-[]-(g:Genus)-[]-(s:SpecificEpithet)-[]-
(o:Occurrence)-[]- (geo:GeopoliticalUnit)-[:sameAs]->(l:Location)-[] ->(c:Climate)

RETURN c.clima_tp_umidade, c.clima_zona, c.clima_distr_umid, c.clima_temperatur

c.clima_tp_umidade	c.clima_zona	c.clima_distr_umid	c.clima_temperatur
"semi-úmido"	"Tropical Brasil Central"	"4 a 5 meses secos"	"quente - média > 18° C em todos os meses"
"semi-úmido"	"Tropical Brasil Central"	"4 a 5 meses secos"	"mesotérmico brando - média entre 10 e 15° C"
"semi-úmido"	"Tropical Brasil Central"	"4 a 5 meses secos"	"subquente - média entre 15 e 18 ° em pelo menos 1 mês"
"semi-úmido"	"Tropical Brasil Central"	"4 a 5 meses secos"	"quente - média > 18° C em todos os meses"
"semi-úmido"			
"semi-úmido"			
"semi-úmido"			
"semi-úmido"			
"semi-úmido"			
"semi-úmido"			
"semi-úmido"			
"semi-úmido"			
"semi-úmido"			

Started streaming 3535 re

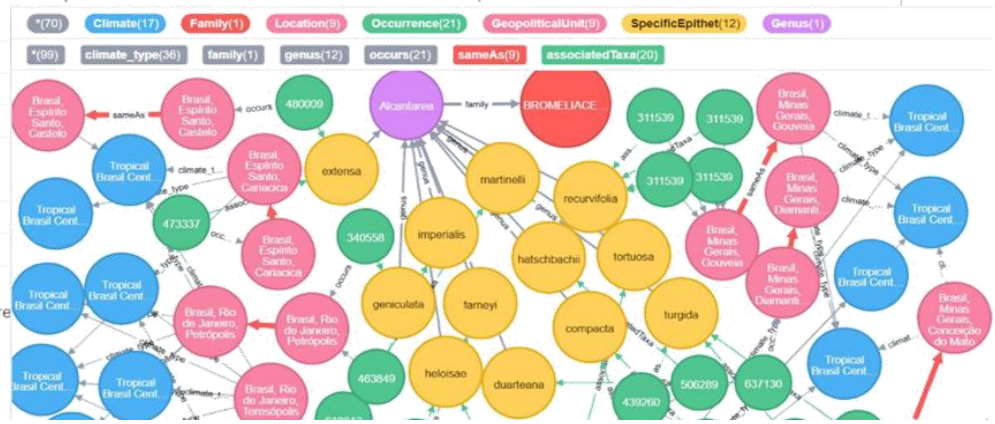


Fig. 5. Climate variations for a given family. Note that the red arrows are the bridges between both datasets

```
MATCH (f:Family)-[]-(g:Genus)-[]-(s:SpecificEpithet)-[]-(o:Occurrence)-[]-
      (geo:GeopoliticalUnit)-[:sameAs]->(l:Location)-[]->(v:Vegetation), (l:Location)-
      []->(b:Biome)
WITH f,v,b, count(o) as Total WHERE Total >= 200
RETURN DISTINCT f.family, b.biome, v.vegetation, Total ORDER BY Total desc
```

f.family	b.biome	v.vegetacao_radam_nmuveg	Total
ASTERACEAE	Cerrado	Savana Arborizada com floresta-de-galeria	680
ASTERACEAE	Cerrado	Savana Arborizada sem floresta-de-galeria	623
ASTERACEAE	Mata Atlântica	Floresta Ombrófila Densa Montana	452
ASTERACEAE	Cerrado	Savana Parque com floresta-de-galeria	393
BROMELIACEAE	Mata Atlântica	Floresta Ombrófila Densa Montana	379
ASTERACEAE	Cerrado	Savana Florestada	333
ASTERACEAE	Cerrado	Savana Parque sem floresta-de-galeria	327
ASTERACEAE	Mata Atlântica	Savana Gramíneo-Lenhosa sem floresta-de-galeria	321
ASTERACEAE	Cerrado	Savana Gramíneo-Lenhosa sem floresta-de-galeria	319
ASTERACEAE	Cerrado	Savana Arborizada	315
ASTERACEAE	Cerrado	Savana Arborizada sem floresta-de-galeria	310
ASTERACEAE	Caatinga	Savana Arborizada sem floresta-de-galeria	280
ASTERACEAE	Amazônia	Savana Arborizada com floresta-de-galeria	278
ASTERACEAE	Mata Atlântica	Savana Arborizada	271
ASTERACEAE	Mata Atlântica	Floresta Ombrófila Densa Submontana	267
BROMELIACEAE	Mata Atlântica	Floresta Ombrófila Densa Submontana	264

Started streaming 22 records after 1301 ms and completed after 1301 ms.

Fig. 6. Biomes and vegetation type of the most frequent families (≥ 200)

5 Conclusion

This work describes the modeling of JabotG, the first version of an RDF graph that will make available in the Web of data, data about the main collection of Jabot database, which is maintained by IPJBRJ. It reports on the use of standard vocabularies and unique identifiers, which facilitates not only dataset integration but also navigation and query over such data. The IBGE/INDE dataset is also described, as an under construction dataset. Although this dataset is now integrated locally to JabotG, the idea is that it must be hosted and maintained separately. Once both datasets are available publicly, it is possible to query and navigate through them, as illustrated by some of the queries.

JabotG is now in its very first version and it is already available as a private endpoint. It means it is internally available as an RDF graph and that it may be queried (SPARQL queries), downloaded or referred to by other local datasets. In a near future, it will be publicly available on the Web of Data. It has been under construction, and as an ongoing work, we are now focusing on reusing geonames vocabulary. In addition, next versions will include other Jabot collections. Finally, JabotG publication process is now under systematization in order to facilitate the maintenance of updated data.

References

1. Brazilian Flora 2020 under construction, http://floradobrasil.jbrj.gov.br/
2. The herbarium catalogue - royal botanic gardens, kew. Published on the Internet; http://www.kew.org/ herbcat, accessed: 2017-07-10
3. The plant list (2013). version 1.1. Published on the Internet; http://www.theplantlist.org/, accessed: 2017-07-10
4. Tropicos - missouri botanical garden. Published on the Internet; http://www.tropicos.org/, accessed: 2017-07-10
5. Baskauf, S., Wieczorek, J., Deck, J., Webb, C., Morris, P.J., Schildhauer, M.: Darwin core rdf guide. Biodiversity Information Standards (03 2015), http://rs.tdwg.org/dwc/terms/guides/rdf/
6. Baskauf, S.J., Wieczorek, J., Deck, J., Webb, C.O.: Lessons learned from adapting the Darwin Core vo-cabulary standard for use in RDF. Semantic Web 7(6), 617–627 (2016)
7. Bizer, C., Heath, T., Berners-Lee, T.: Linked data - the story so far. Int. J. on Semantic Web and Infor-mation Systems 5(3), 1–22 (2009)
8. Cooper, L., Meier, A., Elser, J.L., Preece, J., Xu, X., Kitchen, R.S., Qu, B., Zhang, E., Todorovic, S., Jaiswal, P., Laporte, M.A., Arnaud, E., Carbon, S., Mungall, C., Smith, B., Gkoutos, G., Doonan, J.: The Planteome Project https://pdfs.semanticscholar.org/f944/2886a776b22a5dc6d1ce1955c0785c4f7e4d.pdf
9. Graham, C.H., Ferrier, S., Huettman, F., Moritz, C., Peterson, A.T.: New developments in museum-based informatics and applications in biodiversity analysis. Trends in Ecology and Evolution 19(9), 497–503 (sep 2004), http://linkinghub.elsevier.com/retrieve/pii/S0169534704002034
10. Jaiswal, P., Avraham, S., Ilic, K., Kellogg, E.A., McCouch, S., Pujar, A., Reiser, L., Rhee, S.Y., Sachs, M.M., Schaeffer, M., Stein, L., Stevens, P., Vincent, L., Ware, D., Zapata, F.: Plant Ontology (PO): A controlled vocabulary of plant structures and growth stages. Comparative and Functional Genomics (2005)
11. Silva, L.A.E., Fraga, C.N., Almeida, T.M.H., Gonzalez, M., Lima, R.O., Rocha, M.S., Bellon, E., Ribeiro, R.S., Oliveira, F.A., Clemente, L.S., Magdalena, U.R., von Sohsten Medeiros, E., Forzza,

R.C.: Jabot botanical collections management system: the experience of a decade of development and advances. Rodriguésia 68(2), 391–410 (2017)

12. Telenius, A.: Biodiversity information goes public: GBIF at your service. Nordic Journal of Botany 29(3), 378–381 (2011)

13. Tygel, A., Auer, S., Debattista, J., Orlandi, F., Campos, M.L.M.: Towards cleaning-up open data portals: A metadata reconciliation approach. In: Tenth IEEE International Conference on Semantic Computing, ICSC 2016, Laguna Hills, CA, USA, February 4-6, 2016. pp. 71–78. IEEE Computer Society (2016), https://doi.org/10.1109/ICSC.2016.54

14. Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., Robertson, T., Vieglais, D.: Darwin core: An evolving community-developed biodiversity data standard. PLOS ONE 7(1), 1–8 (01 2012), https://doi.org/10.1371/journal.pone.0029715

Case Study: Ontological Model to Categorize, Enrich and Allow Open Access to Bibliographic Data

Ana María Fermoso García[1], María Isabel Manzano García[1], and Carlos Hernández Tamayo[1]

[1] Pontifical University of Salamanca, Compañia 5, 37002 Salamanca, Spain
afermosoga@upsa.es; mmanzanoga@upsa.es; kambuaset@hotmail.com

Abstract: This article proposes an ontological model based on the design of two ontologies to publish the bibliographic records of a university library as Linked Open Data. We will also demonstrate how open publication allows the data to be linked or enriched with other external sources, primarily SKOS. The practical aspect of the proposed ontological model is rooted in the OpenBiblioUPSA software system, which is presented as a case study. The ontological model is based on two ontologies. The first is based on LD4L, a bibliographic ontology supported by reference bodies such as Cornell University Library, the Library Innovation Lab at Harvard University, Stanford Libraries, and with the participation of the Library of Congress. The second ontology is unique to our university and serves to classify university library resources by category and subject area. Using both ontologies, the system allows the open publication and categorization of resources based on their MARCXML record. Data queries may also be made according to various criteria and enriched with additional external sources. Finally, the system will provide access to other sources using a SPARQL endpoint.

Keywords: Linked Open Data for Libraries (LD4L), Datacite, SKOS.

1 Introduction

The aim of this work is to propose an ontological model for the publication and categorization of bibliographic data as open data. This model will allow the data to be linked or enriched with other external sources, primarily in SKOS format, and converted into a data source to which other data can be linked.

The proposed ontological model has been validated through a case study or a real software system currently in use in our University library: the OpenBiblio project.
Different strategies can be used to make and publish the data in open format. The most effective is Linked Open Data (LOD) which is known as a 5 star format proposed by Tim Berners-Lee [1] for open data publication.

Making data open is a sign of transparency, which is increasingly solicited among institutions. In fact, libraries have spent centuries promoting the use of open access, transparency and accuracy [4] and have begun to take steps in the area of linked open data [5][6][7][8]. This mode of publishing information allows for the information itself to be enriched with external sources as well as serve as a source to others, and link all published information on any subject matter. Due to these characteristics, LOD technology will also optimize and facilitate the work of librarians [3]. Moreover, European and national legislation has reiterated the importance of ensuring that the data and metadata from institutions such as libraries also begin to form part of the universe of open data, facilitating their search and reutilization [3]. On the other hand, it is necessary to underscore how important University libraries are becoming in terms of their support of research within institutions, which includes research data within the spectrum of data to be processed and linked in the university LOD.

The case study presented and implemented in the OpenBiblio software system uses two ontologies designed for our system which allow any bibliographic record from the university library catalogue to be published as linked open data using the LD4L model *(https://www.ld4l.org/)*. Additionally, the records can be classified according to the subject area or topic addressed. Finally, the starting information for any record in our library can be enhanced by information from external sources, if

available, primarily SKOS sources associated to the subject area to which the record pertains, such as the list of subject headings (LEM) from the Ministry of Culture *(http://id.sgcb.mcu.es/),* or the Universal Decimal Code (UDC) (*http://www.udcc.org/)*, the format used in the library when establishing a signature for locating resources.

Library records may also be accessed from external sources using a SPARQL endpoint, also implemented in the case study.
In the subsequent sections of this article, we will analyze the different concepts used in our research and later focus on the ontological models utilized as the basis of our system. Finally, we will conclude with an analysis of the contributions of our system.

2 Concepts and Definitions

The previous section has already mentioned several concepts whose meaning should be clarified, as they form the basis of our proposal. This section will refer to each of these concepts.

2.1 Linked Open Data using LD4L and DataCite

Linked Open Data (LOD) *(http://linkeddata.org/)*, is a format, or better expressed, a set of standards for the most efficient means to publish open data, focusing specifically on the use of semantic technology. To this end, and by applying norms and standards such as RDF *(https://www.w3.org/RDF/)* or OWL, *(https://www.w3.org/OWL/),* the data are interoperable and can be interlinked and enhanced through the Web, which itself is conceived as a set of interlinked concepts and not single pages connected only through explicit web links. This is known as Tim Berners-Lee's 5-star open data plan [1].

In our particular case, we have used Linked Data for Libraries (LD4L) *(https://www.ld4l.org/)* as both the semantic and bibliographic format to publish our data.

Linked Data for Libraries (LD4L) is a project carried out among key institutions in the community of library services, specifically Cornell University Library, the Library Innovation Lab at Harvard University, Stanford Libraries, and with the participation of the Library of Congress. The aim of the project is to facilitate searches for academic information in particular and bibliographic information in general among these universities. LD4L is a format of linked data composed of bibliographic and biographic descriptions of persons and entities. Using such a format, data could subsequently be enriched and linked with data from other sources, primarily other libraries, to complete the information.

LD4L has the same nucleus as Bibframe, but enhances the ontology by expanding to others such as VIVO or FOAF (*http://xmlns.com/foaf/spec/).* The latter complements the former by allowing for the addition of information regarding the identification of academic organizations, their members and statements. The combination resulted in LD4L, a more complete ontology which serves as an exclusively bibliographic cataloguing standard.

Given how complete LD4L can be to format university library resources and expertise, in addition to other important aspects related to university activities, and bearing in mind the prestigious institutions it comprises, we have selected LD4L as the model to shape our own experts and their works.
On the other hand, *DataCite (https://www.datacite.org/)* is a scheme of metadata to publish research data. It was created in 2009 with the primary aim of facilitating access to research data through the Internet, and has expanded throughout Europe, North America, Asia and Australia.

It is a relatively easy model which stands out for including concepts such as ORCID, *(https://orcid.org/),* the fundamental identifier used by researchers, as well as information relevant to grants and funding for the support of investigators and their research.

In our case, we use DataCite, whose semantic scheme is shown to complement LD4L, adding aspects such as the ORCID associated with a researcher.

Finally, the library currently contains all of its records in MARC21 format (*https://www.loc.gov/marc/bibliographic/*), a preeminent format for bibliographic archives. This standard also has a corresponding equivalent in XML known as MARCXML (*www.loc.gov/standards/marcxml/*).

The idea is to obtain information on our records in this standard and convert them to a new semantic format derived from LD4L and DataCite.

2.2 Libraries classification systems

All libraries must have a classification system and one of the most recognized worldwide is the Universal Decimal Classification (UDC) (*http://www.udcc.org/),* which has been managed by the International Federation for Information and Documentation (FID) since 1992.

This is a universal classification scheme, and the most important in the world for all fields of knowledge, particularly used in bibliographic services, documentation centers and libraries around the world. Its use facilitates the process of indexing and recovering bibliographic resources.
These advantages have induced the use of UDC as the classification scheme in our own library, with minor adaptations to our specific needs and areas of knowledge. In this regard, particularly in the field of ecclesiastical sciences, for which our University has long been considered as a benchmark, this classification scheme will be the basis for a possible ontological model in the field, which will be subsequently discussed as one of the key contributions of our project.

On the other hand, the list of subject headings, or LEM *(http:///id.sgcb.mcu.es/)*, promoted by the Ministry of Education and Culture and used primarily in public libraries, is basically a categorization system that references the subject area related to the content of a bibliographic resource or work. It is a more specific means of classifying a work not just with regard to the field to which it belongs, but also the specific subject matter in question.

An additional advantage of LEM and UDC is that both have already been adapted to a semantic format through SKOS. SKOS *(https://www.w3.org/TR/SKOS-reference/)* is a W3C recommendation that uses a semantic format built upon RDF and OWL, designed to precisely represent the content of conceptual schemes such as subject heading systems, thesauri or taxonomies.

This allows us to easily link and enrich our specific resources and their respective topics with the LEM and UDC material available in SKOS format. This complies with one of the objectives of web semantics in general and LOD in particular: the enrichment of data by linking them, through semantic structures, with other related data.

3 Ontological Model

The main feature of our OpenBiblio system, currently in use in our University library, but also exportable to other similar types of libraries, is its ontological model, which permits, among other things, the publication of bibliographic records in Linked Open Data format, as well as the possibility to link and enrich these data with SKOS sources.

In this project, the basis of our system is represented by two ontologies specifically designed to catalogue and classify data according to theme or bibliographic information.

On the one hand, we have a version of the LD4L ontology which has been adapted to our needs, supported by notable university libraries. On the other hand we have our own ontology specially designed for thematic classification according to the content of our bibliographic resources. This ontology has the additional advantage of dynamic adaptation, allowing each university to define its own hierarchy or list of topics.

In the following sections, we will detail the characteristics of both ontologies and explain their usefulness. It will be also explained how is possible to convert the bibliographic resources in MARCXML format records, to the new ontologies, that is, to a LOD format. This will allow the information contained in the bibliographic catalogue records to be linked or enriched with data from other sources such as SKOS or DBPedia. In the following paragraphs we will first look at the foundation of each of our ontologies and then explain how they can be enriched by external sources.

3.1 LD4L_UPSA Ontology

As previously mentioned, LD4L is an ontology to catalogue bibliographic records in semantic format, supported by highly regarded institutions, and therefore important as a semantic format to catalogue bibliographic resources. The new format improves existent models such as the BIBFRAME format used in the Library of Congress, which the new model also supports.

The general structure of LD4L is shown in Figure 1.

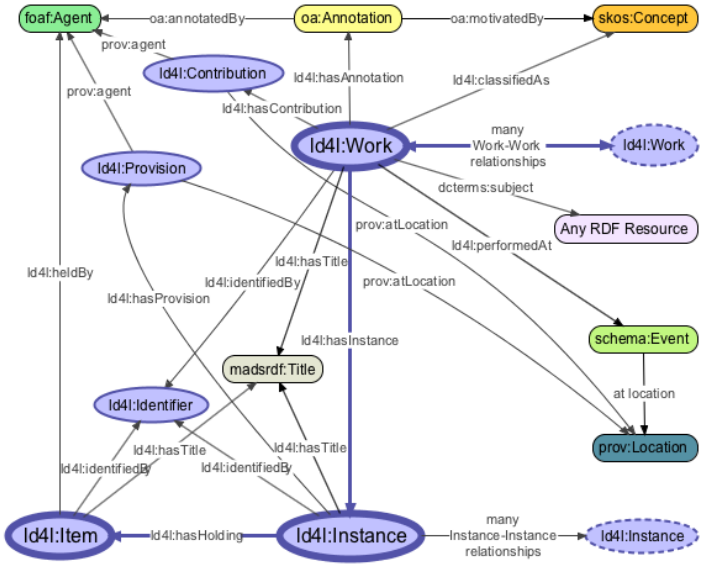


Fig. 1 Ontological Model LD4L

In our ontology, which is fully based on LD4L, we have used the entities or classes of the ontology to classify and catalogue our books and authors. However, we have also enriched the ontology, thus creating our own ontology which we will now refer to as LD4L_UPSA, derived and adapted to our needs from LD4L.

Specifically to the existing Identifier class we have added the following classes: *ISNI (http://www.isni.org/), VIAF (https://viaf.org/)* and most importantly, ORCID. These identifiers make it possible to associate a work with its authors using the *Agent* and *Contribution* classes. The ORCID, or Open Researcher and Contributor ID, can associate the author or researcher to their research or CV, in turn following the DataCite recommendations for facilitating the location and citation of the resources referenced in the research.

Once our LD4L_UPSA ontology is designed, so that it is possible to publish as linked open data, it is critical to determine the model to convert the traditional standard bibliographic format in which the MARCXML bibliographic records are found, to the LD4L_UPSA format.

This involves establishing the correspondence or mapping between the MARCXML fields that contain the information of interest, and our own ontology. Table 1 and Table 2 represent the basis of our

conversion model for publishing the information from our bibliographic and author resource catalogue in Open Linked Data format.

Additionally, new *ObjectProperties* have been added to the ontology entities *Author*: *hasOrcid, hasIsni, hasViaf, hasBirthDate, BornIn,* and *Book* (work in MARCXML): *hasIsbn, WirteIn, hasAuthorName, hasTitle.* Their content will be obtained from the MARCXML associated record.

Table 1. Conversion model from "Authority" class to LD4L_UPSA

MARCXML	LD4L_UPSA
100ª	Instance
24A-viaf	Viaf
24A-orcid	Orcid
24ª-isni	Isni
100D	Date
370ª	Place

Table 2. Conversion model from MARCXML "book" to LD4L_UPSA

MARCXML	LD4L
245ª	Instance,Work,CoverTitle
20ª	Isbn10
100ª	Author
260ª	Place

3.2 BiblioOntology Ontology

The second part or main subsystem of our system permits us to thematically classify our resources. This allows making more specific queries on our data; for example, to know which works on a specific subject or topic are found in our library. This information will be newly obtained from the MARCXML records, which includes the category used by the university itself to categorize its resources, as well as the subject area in which they are located.
However, there are specific areas in which our university is considered a benchmark: ecclesiastical sciences. In this case, the university establishes its own classification thesaurus with the category and signature associated with each one of them.

Each work may also contain references to various subject matters, some of which are included in LEM, LCSH or RAMEAU, and these data are also associated with the work in the new ontology.
In short, the aim is to design a new ontology which allows us to categorize our bibliographic works by theme or subject area. Once the catalogue data has been converted to the new ontology, our system allows for the enrichment of the data from external sources, for example with equivalent topics and categories which exist in semantic models (SKOS) on recognized lists of subject areas, such as those previously indicated or others that may be linked in the future.
The structure of our ontology to categorize our works, bearing in mind the comments we have made, is shown in Figure 2.

There are four main classes: Book, Category, Identifier and Topic, which are in turn interrelated, primarily through the Book class, which is associated with identifiers as well as subject areas and categories that determine the type of content.

In order to associate these characteristics to a book, a series of ObjectProperties will be defined in the book (*Libro*) class, such as *hasCategory, hasCategoryBook, hasEquivalentSubject, hasSubjectBook, has IDcategory, hasIDDocument (hasISBN,hasOCLC).* These properties allow to define class relations; for example, a book with the category or subject area to which it pertains.

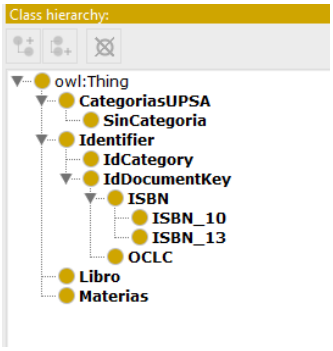


Fig. 2. BiblioOntology class hierarchy

Finally, the ontology allows to classify any work or book from the MARCXML fields such as: title, book identifier, UPSA category, and additionally, if they exist, the subject area or areas associated to the work, to create a instance in BiblioOntologyUPSA.

4 Linking UPSA resources with external sources

The advantage of converting our catalogue of bibliographic records to a Linked Open Data format and using semantic technologies to link to external sources, is that our data will be enriched or linked with data from external sources. Specifically it will be possible to link data with DBPedia as well as LEM using the SKOS format.

Once our resources are converted to our LD4L_UPSA format, they can be queried by author or work. If any existent information about the author appears in DBPedia (*http://es.dbpedia.org/)* or Wikipedia (*https://es.wikipedia.org),* it will be displayed in semantic format. Similarly, their research data will be linked to the information using the ORCID identifier for each author.
In the same way, for each resource found, it is possible to expand information on the category and subject area of this resource data in SKOS format, from the LEM list of subject headings from the Ministry of Culture, or from any other that may exist in LOD format for the category and/or subject of the resource.

5 Conclusions

As the final conclusion for this paper, we would like to highlight the main contributions of our system as a means of formatting information into semantic format, which additionally allows enriching the information and/or serving as a source of data for other external data sources.

We have designed two ontologies which allow semantically modeling works and categories in the world of university libraries. The first ontology uses the LD4L as a basis and adapts it with minor changes. The second ontology was uniquely designed and permits the semantic categorization of bibliographic resources according to the topic or subject area of the content of the resource itself.
We have also implemented the system or application online (*http://dataupsa.upsa.es/BiblioOntologyUPSA/*), making it possible to take advantage of these ontologies.

The web application includes a query system for bibliographic resources by book, author, subject area (category), and it have even added a visual representation of the query. The query results can be

enriched with external information from other LOD sources, mainly, but not only, in SKOS format. Finally, the online service also provides a SPARQL endpoint to allow be queried from external sources. In short, we have a new model to format information which uses semantic technologies within the field of libraries, allows publishing its content in Linked Open data, and also enriches the data with external sources in SKOS format or DBPedia, while allowing the data to be linkable from other external sources using an own SPARQL Point.

References

1. Berners-Lee, Tim “Linked data. Design issues”, http://www.w3.org/DesignIssues/LinkedData.html (2006)

2. Fermoso-García, Ana; Manzano-García, María-Isabel; Armero-López, Álvaro; Hernández-Hernández, Álvaro “Apertura y publicación de datos bibliográficos con linked open data. Un caso práctico”.En: XV Workshop Rebiun de Proyectos Digitales "Datos y Bibliotecas”. http://hdl.handle.net/10234/163387 (2016)

3. Giusti-Serra, Liliana; Santarém-Segundo, José-Eduardo. “Ó catálogo da biblioteca o linked data”. Em Questao, v. 23, n. 2, p. 167-185 http://dx.doi.org/10.19132/1808-5245232.167-185. (2016)

4. Hallo, María; Luján-Mora, Sergio; Maté, Alejandro; Trujillo, Juan. 2016, "Current state of Linked Data in digital libraries", Journal of Information Science, v. 42, n. 2, p. 117.

5. Seikel, Jones; Seikel Michele. “Linked Data for Cultural Heritage”. American Library Association Collections and Technical Services. Chicago (2016)

6. Sulé, Andreu; Centelles, Miquel; Franganillo, Jorge; Gascón, Jesús. "Aplicación del modelo de datos RDF en las colecciones digitales de bibliotecas, archivos y museos de España." Revista española de Documentación Científica. v 39, n 1: e121 http://dx.doi.org/10.3989/redc.2016.1.1268 (2016)

7. Tharani, Karim. "Linked Data in Libraries: A Case Study of Harvesting and Sharing Bibliographic Metadata with BIBFRAME", Information Technology & Libraries, v. 34, n. 1, p. 5-19. (2015)

8. Torre-Bastida, Ana-Isabel; González-Rodríguez, Marta; Villar-Rodríguez, Esther “Datos abiertos en lazados (LOD) y su implantación en bibliotecas: iniciativas y tecnologías”. El Profesional de la Información, n 4, p. 113-120. https://doi.org/10.3145/epi.2015.mar.04 (2015)

Data Model of the STKOS Metathesaurus

Zheng Liu[1], Shanshan Ji[1], Wen Song[1], and Tan Sun[2]

[1] National Science Library, Chinese Academy of Sciences, Beijing 100190, China
{liuz, jishanshan, songw}@mail.las.ac.cn
[2] Agricultural Information Institute, Chinese Academy of Agricultural Sciences, Beijing 100081, China
{suntan@caas.cn}

Abstract. To achieve better knowledge organization and semantic linking of the scientific and technological information resources, STKOS project constructs a knowledge organization system called Metathesaurus, which is a compilation of concepts, terms, relationships, and associated information from a variety of existing controlled vocabularies in scientific and technological area. The fundamental unit of STKOS Metathesaurus is the concept, and each concept may link to more than one terms from various vocabularies. All the concepts and terms form a knowledge network which can be used to locate, describe, and organize knowledge implied in digital resources. This paper presents the framework and data model of STKOS Metathesaurus, explicates how to standardize the concept expressions extracted from different sources, and link the alternative names of the same concept together.

Keywords: knowledge organization, data model, STKOS Metathesaurus.

1 Introduction

To help user find knowledge and relationship of information form vast amounts of English Scientific & Technological Literature collection, improve the capability to all kinds of services based on Literature, National Science and Technology Library (NSTL) in China has initiated a program called "Development and Application of Knowledge Organization System for Foreign Scientific & Technological Literature" (STKOS), which is granted by the Ministry of Science and Technology of China. The purpose of STKOS project is to facilitate computer systems to "understand" the language of Science and Technology. STKOS project works out and distributes the Scientific & Technological Knowledge Organization System and assisted software, tools (programs) for system developers' use in building or enhancing electronic information systems that create, process, retrieve, integrate, and/or aggregate science and technology data and information. [1, 2]

The fundamental work of STKOS is constructing the Scientific & Technological Knowledge Organization System which we call it Metathesaurus. STKOS Metathesaurus builds a network of concepts, terms, relationships, and associated information from a variety of existing controlled vocabularies and keywords in scientific and technological area. In the entire knowledge system, the core units are concepts and the basic units are terms. Terminologies and concepts are connected by synonymous relations, while the original relations of source vocabularies are also retained. [3] STKOS metathesaurus can serve as a knowledge source for developers of scientific and technological information applications, and can be used as a tool in knowledge management or retrieval systems.

As the basis of Metathesaurus construction, we define a data model to explicitly describe the relationships between standard concepts and source terms, and presents a framework supporting the construction process. If different vocabularies use different terms for the same concept, or if they use the same term for different concepts, then this will be distinguished and represented in the Metathesaurus. This paper gives a detailed introduction of the data model and framework.

2 STKOS Metathesaurus Data Model

STKOS Metathesaurus integrates various knowledge organization systems such as glossaries, classifications, and thesauri. These sources are always multidisciplinary, multi-format and multilingual. Thus, the key point is to standardize the concept expressions extracted from different sources, link the alternative names of the same concept together, and make sure original relationships between concepts can be tracked.

To design a rational model, we do a deep research on current KOS data models. Take the developments of thesaurus standards as example, there are two types of data model: term-centric and concept-centric. Term-centric data model includes preferred terms and non-preferred terms, and the relationships are built among terms. Concept-centric data model such as ISO 25964-1 distinguishes clearly between concepts and terms. In ISO 25964-1 standard, the tags BT, NT and RT are retained (because these have been widely used in thousands of existing thesauri), but clarifies that the relationships they indicate are between concepts, not terms. [4] We finally choose concept-centric data model to construct STKOS Metathesaurus, and according to the demands from the scientific and technological area, we do some changes and improvements based on the data models of ISO 25964-1. [5]

The data model of STKOS Metathesaurus contains six core elements: source vocabulary, source term, scientific and technological term, standard concept, category and category class. Table 1 lists the definitions of the core elements, and Fig.2 briefly illustrates the properties and relationships among them.

Table 1. Definition of the STKOS Metathesaurus Element

Element	**10. Tag**	**11. Definition**
Stkos:Vocabulary	Source Vocabulary	14. Existing comprehensive or specialized knowledge organization systems, including thesauri, subject headings, glossaries, classifications, dictionaries. These are the basic elements of STKOS Metathesaurus.
Stkos:sourceTerm	Source Term	17. Terms from the source vocabularies which are evaluated, transformed and normalized.
Stkos:STTerm	S&T Term	20. Science and Technology terms which are used to express S&T concepts, and mainly are from the source terms. Terms indicating to one concept may have various forms. Through recognizing homographs and synonyms, we can link terms with the same meaning together, choose one and support for future concept-based search.
Stkos:Concept	STKOS Standard Concept	23. A concept is an abstract idea representing the type and fundamental characteristics of some set of objects. The label of STKOS standard concepts derive from the S&T terms, and are defined by statistical analysis and expert checking.
Stkos:Category	Category	26. The collection of category classes, usually in classification structure, can be used to organize the concepts or terms.
Stkos:CategoryClass	Category Class	29. Classes in the category, generally denotes a specific subject area.

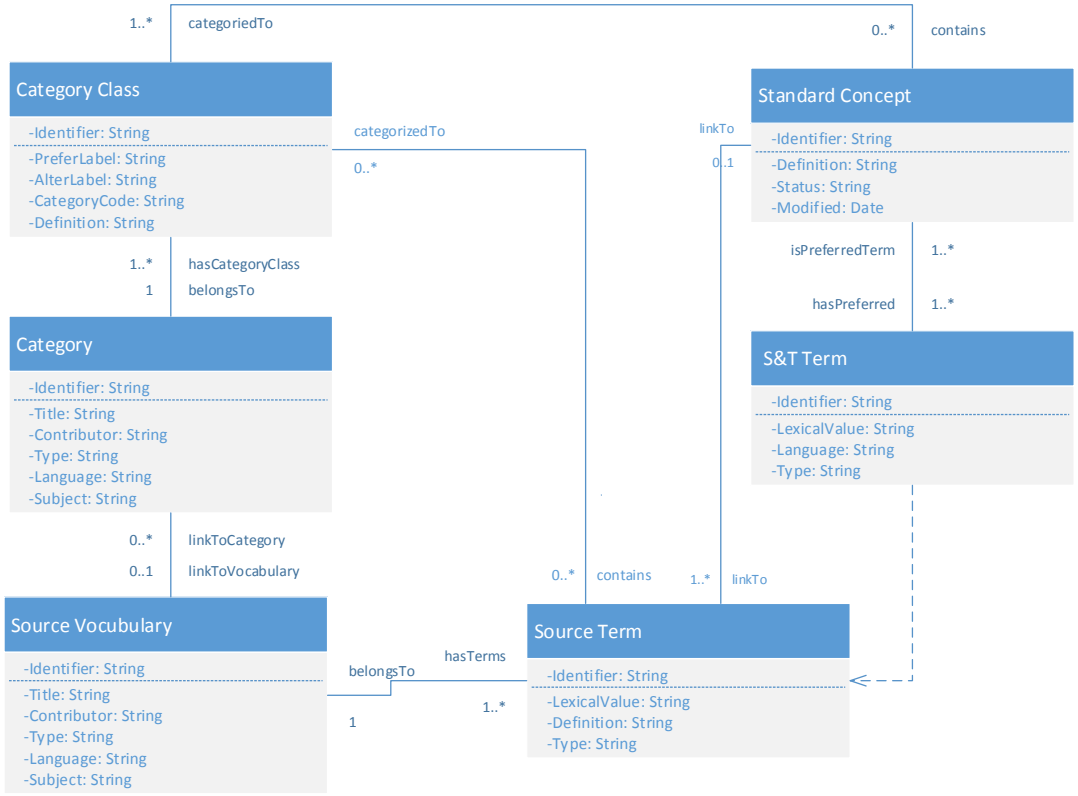


Fig.2. STKOS Metathesaurus Data Model

3 STKOS Metathesaurus Framework

To implement the construction of STKOS Metathesaurus, we design a new kind of technical framework based on existing researches, such as UMLS[6], AGROVOC[7], STERNA[8], and NeON[9].

The framework of STKOS Metathesaurus is comprised by five parts: KOS Registration System, Source Vocabulary Warehouse, Term Base, Standardized Concepts, and Category Systems.

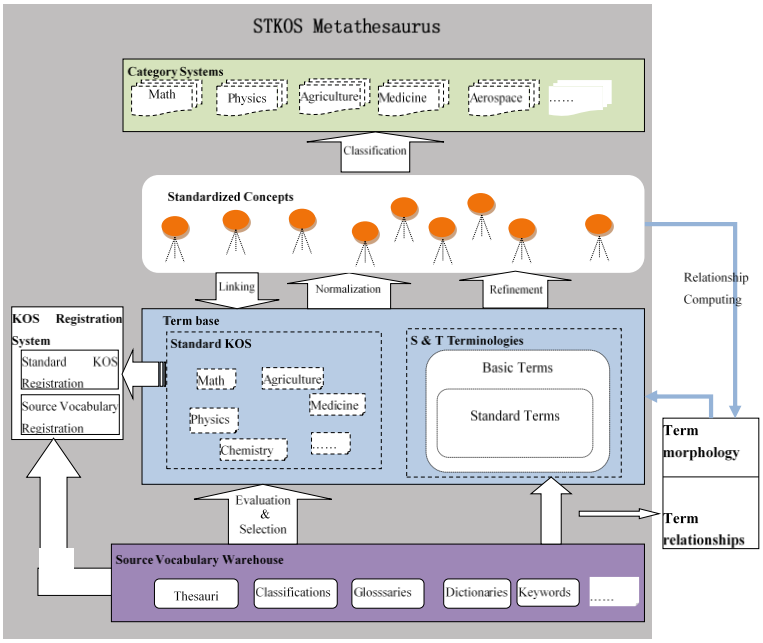


Fig.1. STKOS Metathesaurus Framework

- **KOS Registration System**: Description and registration for source vocabulary and standard Knowledge Organization Systems (KOS).
- **Source Vocabulary Warehouse**: The main sources are existing thesauri, classifications, code sets, and lists of controlled terms used in S & T statistics, indexing and cataloging literature. Through data collection, evaluation, selection, and format conversion, those resources are then uploaded to the vocabulary warehouse.
- **Term Base**: After evaluation of the source vocabularies, highly rated ones are marked as standard knowledge organization systems (KOS). We select terminologies or keywords from this standard KOS, thus form the basic terms which inherit the original term variations and relationships. Through word stemming and lemmatization, related terms which are characterized as lexical variants are grouped into a cluster. Then after semantic merging and disambiguation, terms with the same meaning but may have different expression forms are merged. In cases where a string has multiple meanings, there is a separate term for each distinct meaning.

- Standardized Concepts: Based on the cluster of synonymous terms, one term was chosen as the preferred term representing the synonymous group to form the standard concept. With the assists of computer and domain experts, the preferred terms are translated into Chinese, and every concept is classified and defined. The resulting concept network links alternative names of the same concept together and used to identify semantic relationships between different concepts. Also, it reserves the relationships between concepts and source terms.
- Category Systems: According to discipline classification or hierarchy, the terminologies and standardized concepts and are classified. In future STKOS applications, classification systems can be used for concept organization and automatic clustering.

4 Result

We make a further refinement of the framework and data model, and build a set of collaborative construction platforms to support the building process of STKOS Metathesaurus [3]. After four years' work, STKOS Metathesaurus now contains about 615,384 standard concepts, which can link to 2,321,681 source terms from more than 201 vocabularies. STKOS Metathesaurus can be used in NSTL system or the third party systems, to help user find knowledge and relationship of information from vast amounts of English scientific and technological literature collection.

Fig.3 shows the general information of “Ultrasonic measurement” which is a concept in STKOS Metathesaurus. This figure is a screenshot from the Sharing Portal of the Scientific and Technological Knowledge Organization System [10, 11].

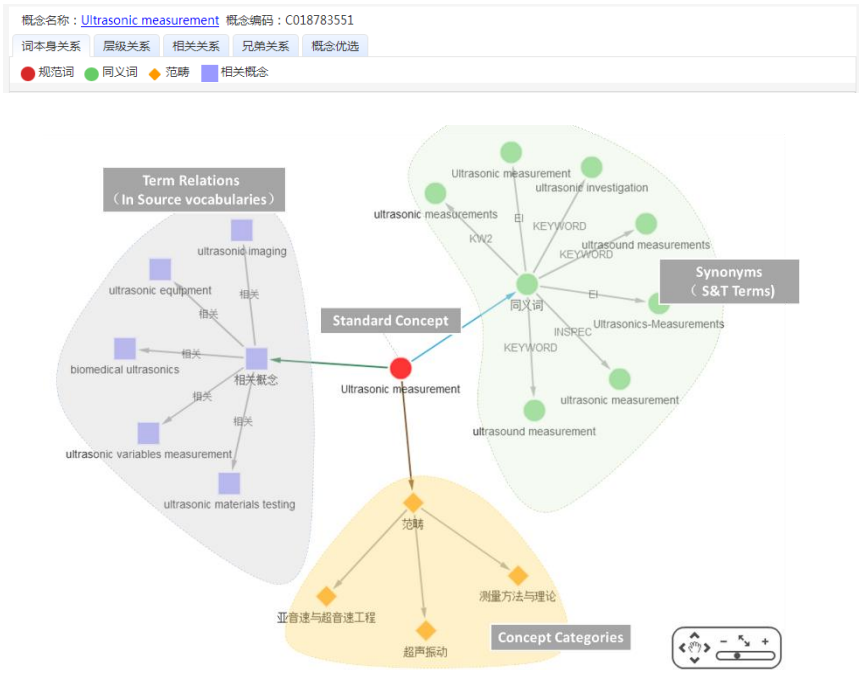


Fig.3. General Information about “Ultrasonic measurement” Concept

5 Future Work

This study constructs a concept-based data model for STKOS Metathesaurus, provides a concept-centered solution to organize terms from different sources. Take STKOS Metathesaurus as an infrastructure, more research applications such as large-scale semantic calculation, intelligent retrieval, automatic processing, and knowledge service can be implemented.

In the future, our research will put more efforts on STKOS data publishing, sharing and linking. On one hand, to fully describe the elements in STKOS Metathesaurus data model, we will make some extensions based on existing SKOS model. Thereby, all STKOS data can be published in well described format. On the other hand, we will develop automatic data linking algorithms to build the relationships between STKOS data and Linked Open Data, thus facilitate the STKOS data sharing.

References

1. Sun T., Liu Z.: Methodology framework of knowledge organization system for scientific & technological literature. Library and Information, 1, 2–7 (2013).

2. Zhao R., Xian G., Kou Y., et al.: Construction of a domain knowledge service system based on the STKOS. Journal of Data and Information Science, 7(3), 50-61 (2015).

3. Sun H., Li D., Li J.: Design and implementation of the STKOS quality-control system. Chinese Journal of Library & Information Science, 8 (3), 38-49 (2015).

4 Dextre Clarke S G, Zeng M L.: From ISO 2788 to ISO 25964: The evolution of thesaurus standards towards interoperability and data modelling. Information Standards Quarterly (ISQ), 24(1), (2012).

5. Iso25964 standards. http://www.niso.org/schemas/iso25964/, last accessed 2017/7/6.

6. Bodenreider O.: The unified medical language system (UMLS): integrating biomedical terminology. Nucleic acids research, 32(suppl_1), D267-D270 (2004).

7. Rajbhandari S, Keizer J.: The AGROVOC concept scheme–a walkthrough. Journal of Integrative Agriculture, 11(5), 694-699(2012).

8. STERNA (Semantic Web-based Thematic European Reference Network Application). https://www.prestocentre.org/resources/projects/semantic-web-based-thematic-european-reference-network-application-sterna, last accessed 2017/7/6.

9. NeOn project. http://www.neon-project.org/nw/Welcome_to_the_NeOn_Project, last accessed 2017/7/6.

10. Wang Y., Zhang Z., Li C., et al.: The Design and Implementation of Open Engine System for Scientific & Technological Knowledge Organization Systems, 31(10), 95-101(2015).

11. Sharing portal of Scientific and Technological Knowledge Organization System. http://stkos.las.ac.cn, last accessed 2017/7/6.

InDisco: Instance Discovery using Mining Algorithms over User Interactions in Semantic Social Networks

Mouaz Alabsawi, Amr Elmougy, and Slim Abdennadher

Computer Science and Engineering Department
German University in Cairo, Egypt
mouaz.esam@guc.edu.eg

Abstract. In general, mining data sets aim to discover new association rules and patterns. Classical data mining techniques utilized tables of data as input, thus retrieving plain associated data. Using semantic data as input would yield an output in the form of semantic results. In this paper, we propose a novel method to discover new instances from semantic data sets given an input ontology. The ontology input decides which resources will be selected. Next, a SPARQL query will be automatically generated. This leads to a result set that goes through FP-Growth mining algorithm. The new associated patterns can build new instances thus enriching the data set in the process. This methodology was experimented over a data set of semantic social network in order to show the effectiveness and uniqueness of the results.

Keywords: semantic web, social network analysis, data mining, big data.

1 Introduction

Since the ways of information representation are evolving, the manipulation tools along with the data analysis are also evolving. Association analysis is one of the key methods to discover rules and patterns within large data sets. Although these rules can help in decision making, the knowledge base itself remains at the same state in terms of size and capabilities. After the introduction of the semantic web by Sir Tim Berners-Lee in 2001 and the growth of initiatives such as Linking Open Data community project, new ways of structuring data are available for testing by data mining algorithms. One important question arises, if the data is really semantically modeled, whether we can extract semantic results and how data mining algorithms with abstract inputs can be applied to structured data.

For example, if a data set contains the historical wars in Europe after the Industrial Revolution where the figures are linked to the data with just single predicates, the question is whether we can extract the complete information about each figure. This will lead to queryable data which enriches the knowledge base in the process. Upon exploiting semantic data sets, we further improve association analysis. In other words, complete instances will be extracted which will enrich the knowledge base according to the ontology inputs.

In February 2004, Stephen Downes and Marco Neumann introduced the term semantic social networks. Downes had a vision of how the Internet will look like with semantic social networks. Semantic social networks make use of the ever growing information provided by the users to the network. The power of the semantics appears when the 'friends' and 'interests' are classes. It may happen that two users have a common interest even though they have never met before [6].

One of the main challenges that face having effective semantic social network is that the famous current social networks are satisfied with their scheme type. In addition, the migration to semantic dataset is expensive from a system architecture and security perspective. For instance, Facebook developed its own Graph database with its Graph API through *https://graph.facebook.com/*. The API results are always in JSON format. However, efforts were made in 2012 to provide the linkage and the semantic translation to Facebook data through Graph API [7].

One of the potential research areas in Semantic web technologies and Linked data is how to get the most out of semantic data. The perspective of the agent differs according to what it needs from the knowledge base. The more perspectives provided, the wider the scope and the more available it is to different types of agents. Recent research encompassed learning ontologies, mapping between different ontologies, annotation by information extraction and duplicate recognition. Semantic data engineering faces technological challenges compared with traditional tabular data. However, generating new objects from multiple domains automatically from one existing knowledge base represented in semantic data is more effective than from knowledge base represented in plain or tabular data due to multiple reasons. Semantic datasets are mostly represented as triple stores which provides the flexibility and simplicity away from the regular constrained tabular data. The flexibility is not just in representing the data but also in applying huge migrations. Reasoning is one of the main examples. The querying may appear as competent factor. However, the ontology and searching algorithms ease the process of generating complicated SPARQL queries between different domains which does not exist in tabular data. The execution of queries in semantic data depends on following the indexes of resources in the triples. On the other side, tabular querying counts on the join operation which is extremely costly. For instance in the generated query of the experiment, there was six classes. If they were represented in tables, the complexity becomes $O(\Pi_{i=1}^{k} n_i)$ where n; represents table i; rows count and k; is the number of tables to be joined.

One of golden features of the semantic web is the expressivity, so sophisticated entities are much easier to be represented in RDF than in tables.

The purpose of this study is to enrich the knowledge base of a semantic dataset such that new instances can be generated. Consequently, agents can perform their SPARQL queries over data about classes that has been never explored. Firstly, the ontology of the new instances is read. This is followed by the generation of a SPARQL query to capture its properties. Next, data mining algorithms are applied on the result set of the SPARQL query. As a result, queryable new instances are added to the knowledge base.

The paper is organized as follows. In Section 2 (InDisco), we will introduce the overall methodological approach for investigating our research problem from the very beginning to the final output. In Section 3, we will show how the methodology is applied on a use case data set (Semantic social network). At the end in Section 4, we highlight the conclusion and the future thoughts of our methodology.

2 InDisco

In order to extract the hidden instances of a certain concept in a large data graph, the architecture follows a methodology that takes a concept description as an input and produces its instances as an output. Every ontology has a set of resources that form its meaning like the wheels and seats for a vehicle. The architecture goes through a set of steps to generate the new individuals. Firstly, it extracts the resources that form the input ontology instance. Secondly, using the data set ontologies, it finds out the pattern that connects the needed resources and hence generates a SPARQL query. Thirdly, a mining algorithm is applied over the result set to get the most common associations. Every common association represents a new set of triples to be added to the instance. The input ontology is the ontology that instances are going to be created accordingly. The existing data set also contains a set of ontologies that represent the included instances. Some of the input ontology resources may intersect with some of the included data set ontologies resources in terms of their type. The relation between the input ontology resources in the data set ontologies resources is called the mining pattern.

2.1 The Mining Pattern

A Mining pattern is subgraph of resources inside the ontologies existing in the knowledge base. It consists of a set of concepts.

The necessity behind the mining pattern is to show the connections between different resources which makes generating the SPARQL query an easy task. For example, if you are looking for a trademark of a wheel and its country of origin in a data set of vehicles, the algorithm is going to find out that a vehicle is a thing with two wheels or more. A vehicle has a manufacturer. Every manufacturer has its country of origin. Therefore, the query is going to ask the data set about *countries and wheels of every vehicle*. Then every country will be associated with the corresponding wheel and end users are now able to look for countries of origin of wheels directly since the ontologies form a graph at the end of the day. The connections between different resources have been extracted using consecutive semantic Breadth-first search algorithms (BFS).

The generating process of the mining pattern is divided into small search problems between concepts in order to connect the resource to each other. It is extracted according to the following algorithm:

```
Algorithm 1: Finding the mining pattern between selected resources
Result: SubGraph containing the links between selected resources
get data set ontologies resources;
get input ontology resources;
while All selected resources not visited do
    if current resource is visited then
        BFS between current resource and the next one;
        Add the output path to the SubGraph;
    else
        Continue;
    End
end
SubGraph is ready;
```

The complexity of the above algorithm is the same as the one for the BFS algorithm O(c.length(V+E)) as it passes by every resource once. However, performance of finding the pattern between ontologies is not always a critical issues because ontologies are usually small in their structure as they describe concepts for instance FOAF ontology has 78 resources. Semantic sensor network(SSN) ontology has 107 resources.

2.2 Generating the Query

For now, the mining pattern is defined with its resources and the links between these resources. In order to extract a result set of the input ontology resources, a query should be issued. Extracting SPARQL query from the mining pattern is a straightforward process since the system is aware of the relation between the resources that are supposed to be queried. Every result in the transaction (result) set from the query is a chance to add more triples and expand the new instances.

2.3 Extracting Transactions and Mining Process

After executing the previous query, the result set will represent a set of associations. Every row will represent what it is called a Transaction. Each association is a group of resources that can be added to the instance. The decision of whether the associated resources are going to be added or not is made according to how much the association is common. The popularity of an association is determined by the threshold of the mining algorithm. A priori and FP-Growth algorithms have been applied. FP-Growth showed high performance in large datasets. However, a priori provides wide range of results.

After applying the mining algorithms, the common associations appear. Triples are added to the instances according to resources in every association.

Since the algorithm keeps the subgraph at the first place, the predicates of new triples can be added.

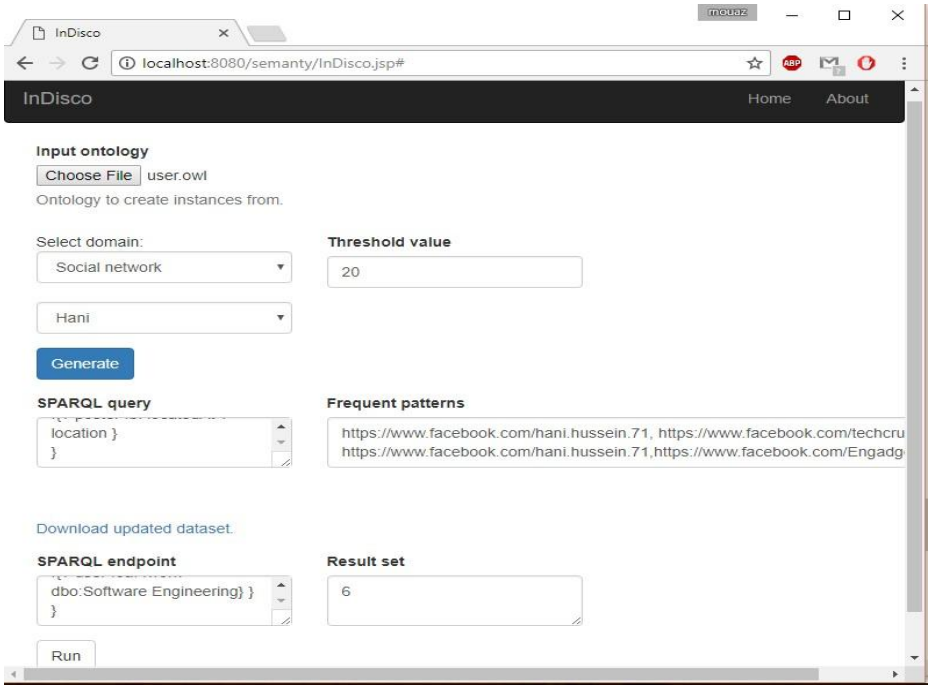


Fig. 1. InDisco in Action

3 Experiment and results

The system has been implemented in Java EE on a local host with Apache Maven. The local host with capacity of 100 GB free on an SSD disk with core i5 processor and 8 GB ram on windows 10 with 64 bit. Jena 3.1.1 library is used for manipulating semantic data sets. The experiment has been held over the interactions on Facebook. The size of participants is 15. 7 participants work in the tech field, 3 participants work in academia, 2 participants work as medical representative, 2 participants work as pharmacist and a dentist. Out of the 15 participants, 9 are males and 6 are females. Their age was between 25 and 59. They approve on providing us the access to their public interactions on Facebook. After applying our methodology with their extracted identities, a survey was conducted to measure the accuracy of the output profiles.

3.1 Ontologies involved

FOAF Friend-of-a-Friend ontology is one of the most popular ontologies in the semantic web world. FOAF initiative is a machine readable ontology to describe people, organizations, groups and some other objects that are related to people. FOAF has its own vocabulary including its classes and properties. FOAF started with describing people. But since people are strongly linked to other things with different properties, FOAF started to describe other activities such as work in organizations, make documents and attend meetings.

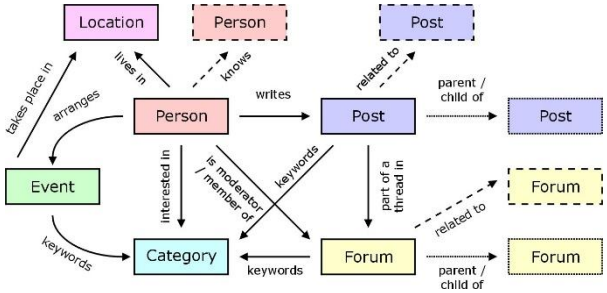


Fig. 2. Small section of FOAF Ontology

Post When a user likes a post, it tells a set of things about him/her. For instance, if he liked a post from Amazon Ireland that Amazon is going to deploy new data center there, so user could be working in the tech field, speaks English and Irish. Or he/she might be just interested in tech news. We cannot decide from one interaction. That is why larger sample size is needed. If he/her liked hundred posts from tech companies in Dublin. The probability of being someone working in the tech field becomes higher. Post object on the graph API does not include everything we need to our experiment directly and fill the post ontology created. Therefore, during the modeling process we used entity extraction API to fullfil post ontology instances. [16]

3.2 FB Graph API

Small app has been created on Facebook to retrieve the information of posts that have been liked by the users. Users gave the app the permission to retrieve the posts they have interacted with.

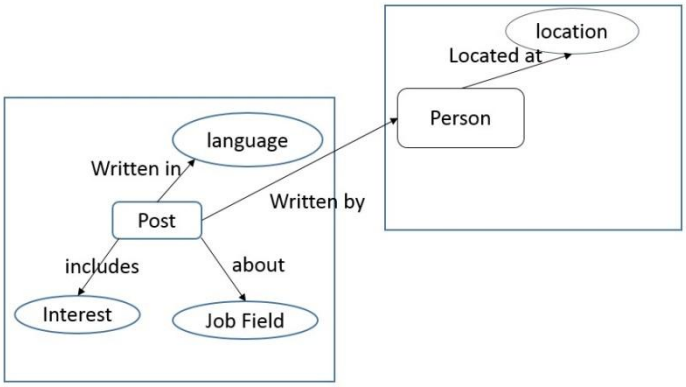


Fig. 3. Post Ontology

3.3 Entity extraction

One of the most essential factors to the experiment is how to parse posts contents on social media to get the entities from rich text online. Web service has been implemented to get the resources of entities in semantic form from dbpedia. The API that is used Dendilion API that is developed to find places, people, brands, and events in documents and social media.[15] Because Dandelion supports a set of languages, any other language that is not supported has been excluded.

3.4 SPARQL generated

Listing 1. The SPARQL query - some variable names changed to look descriptive

```
select distinct ?user ?poster ?location ?interest ?occupation ?language
      where {
{?user a foaf:Person }
.{?poster a owl:Thing}
.{? location a gn:Feature}
.{? interest a fb:Interest}
.{?occupation a fb:Occupation}
.{?language a dbo:language}
.{?post a fb:Post}
.{?user fb:liked ?post}
.{?post fb:writtenBy ?poster}
.{?post fb:about ?occupation}
```

.{?post fb:includes ?interest }
.{?post fb:writtenIn ?language}
.{?poster fb:locatedAt ?location }
}

Pattern	Count
Hani, Business Insider, NYC-US, English, software engineering, technology	72
Hani, Medium, San Francisco - US, English, - freelancing	21
Hani, Asmaa Samir, Cairo - Egypt, English software engineering, technology	27
Hani, Web Course, Cairo - Egypt, English software engineering, technology	25
Hani, National Geographic Adventure, Washington D.C. - US, English -, travel	32
Hani, Sherif AbdEl Aziz, Cairo - Egypt, English -, photography	41

Snapshot of mining results

3.5 Survey questions

A survey has been created to verify the accuracy of the results. Some of the questions results were from scale 1 to 10 such as friends, interests and languages. Because a user can speak one or more language and can be interested in one or more topic. Political and religious views have been excluded upon the request from participants. The rest of questions are yes-no questions.

3.6 Results

The participants were astonished after they saw their new profiles. They were not aware of how much their interactions on social media tell about them and their character. However, the sample result of their interactions on social network did not completely contain the full list of their friends and interests.

Type	Percentage
Interests	88%
Friends	76%
Languages	83%
Job field	93.3%
Location	86%
Overall	84.6%

Accuracy of each result

The accuracy of each result changes from a field to another. Every part was measured separately. Interests, friends and languages were measured on a scale where the rest was measured with yes-no questions.

1. Interests: Interests showed high accuracy since it is closer in the data diagram to the user. There is high probability that a user who like a post about new wedding plans to be interested in wedding planning.
2. Friends: The friends field showed a significant small value due to the interactions we get from users were about public posts. Therefore, the number of extracted entities that are related to users were not enough. Also some social media influencers they use their personal account which made the task harder to differentiate between a close friend and a celebrity because they are the same class.
3. Languages: Languages accuracy had overall average value because some users who follow some pages because its media content not the language. In other words, users are not bounded by the languages they know when it comes to following entities on social network.
4. Job field: Job field surprisingly scored high accuracy with only one participant missed it. After investigation, we found out that the user was overwhelmed with honeymoon plans which made his/her interactions away from its job field.
5. Location: Location also scored a high accuracy. The left 14% in its value was because some users who are much interested in entities from another countries such as US which affected their current location

3.7 Evaluation

Since the methodology made use of different technologies such as data mining and semantic web, so we evaluated the benefit behind each technology. In association analysis, there are three main factors that evaluate the mining process output which are Data source itself, Validity of the results and Interesting- ness of the extracted patterns.

Data resources. When we evaluate the size, structure and context. The size of data was not so large as the first experiment. However, the context of the data (social network interactions) makes the data promising to provide interesting results.

Validity. According to the previous survey conducted to validate the information about the participants, the following chart shows the percentage of each profile accuracy.

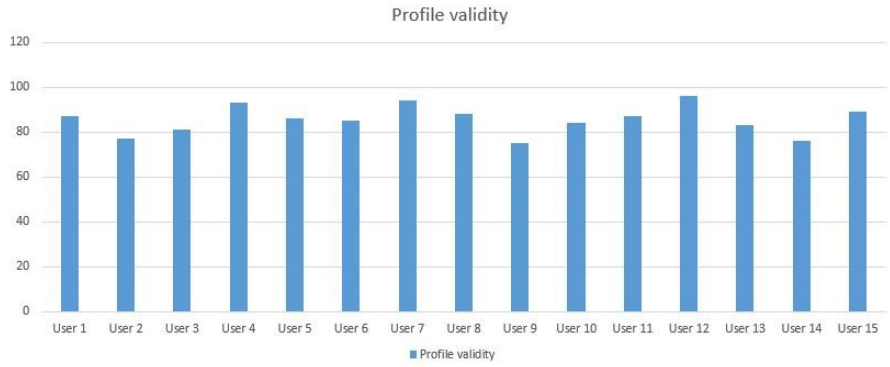


Fig. 4. Validity of each user

Interestingness The main reason behind choosing semantic social network experiment is to show interesting results. Our algorithm produces new instances that can be queried. Therefore, pilot test set of queries were executed to satisfy the motivation behind the experiment.

Listing 2. Query languages known by a certain user

```
select ?language where {
{fb :Marc a foaf:Person }
.{ fb :Marc foaf:knows ?language} }
}
```

Result: [xsd:en-US , xsd:fr-FR]

Listing 3. Number of software engineers in the dataset

```
select count( distinct ?user) where {
{?user a foaf:Person }
.{?user foaf:work dbo:Software Engineering} }
}
```

Result: [6]

On the other hand we conducted the participants to give insights about how interesting they find the results.

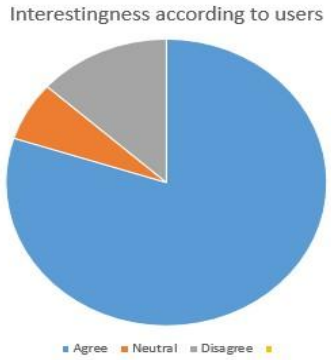


Fig. 5. Overall interestingness of profiles

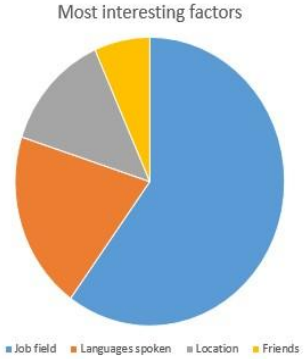


Fig. 6. Most interesting factors

Although the results are promising in the aspect of developing the semantic data, there are a set of limitations in the mechanism. The expansion of ontology individuals is only concerned with classes and objects which means properties are deduced according to the findings of patterns. For instance, if the pattern has a location and a user therefore, the first property (”located at” in our case) between location and user will be picked up. The disadvantages of this approach is that the results of different properties with same class objects become inconsistent.

4 Conclusion and Future Work

In the experimentation, we can show that hidden association patterns are able to construct new relations and statements in order to form new individuals according to the input ontology and a generated SPARQL query between different data set ontologies from different domains. Although it seems that the semantic web is still in the crawling stage. We can see high potential in the direction of improving existing ontologies according to semantic data sets. Improving the ontologies of different contexts will raise the level of knowledge and situation awareness to agents which leads to better understanding and quality decision making.

The implementation of the paper is a combination of semantic web, data mining and social network analysis. The future thoughts can be considered in every of the mentioned directions. In the part of the semantic web, generating SPARQL queries can be more expressive through looking behind classes and types and go to properties. In the aspect of applying data science algorithms over the semantics,

there are still several other algorithms that have not been experimented yet over semantics. When it comes to social network analysis, there is a wide scope of applications. Instances such as communities, spammers and public figures can also be extracted with the same methodology if we have well-structured semantic dataset. Not only social networks can be targeted by our algorithm but also mining linked data in the semantics of IoT in order to give agents better situation understanding and patients medical records to extract diseases description.

References

1. T. Berners-Lee, J. Hendler, & O. Lassila. The Semantic Web. Scientific American, May 2001.
2. Nikos Bikakis, Chrisa Tsinaraki, Nektarios Gioldasis,& Stavros Christodoulakis, XML and Semantic Web W3C Standards Timeline, The XML and Semantic Web Worlds: Technologies, Interoperability and Integration. A Survey of the State of the Art, October 2014.
3. Tim Berners-Lee, Semantic Web on XML, XML 2000 Washington DC, June 2000.
4. RDF Schema 1.1, https://www.w3.org/TR/rdf-schema/
5. OWL Reference, https://www.w3.org/TR/owl-ref/
6. Stephen Downes, The Semantic Social Network., The learning organization 12 (5), 411-417, 2004
7. Facebook linked data via the graph API J Weaver, P Tarjan - Semantic Web, 2013.
8. The influence/impact of Semantic Web technologies on Social Media, Alfonso Infante-Moro, Anca Zavate, Juan-Carlos Infante-Moro, 2015
9. Vrandei, D., Passant, A. & Breslin, G. (2011). Social Semantic Web, Springer-Verlag Berlin Heidelberg. [Accessed on: March 2014].
10. Studer, S., Benjamins, R., & Fensel, D. (1998). Knowledge Engineering: Principles and Methods. Data and Knowledge Engineering, 25,161-197.
11. Finin, T., Ding, L. &Zou, L. (2005). Social Networking on the Semantic Web.
12. Twitterizer, https://github.com/Twitterizer/Twitterizer
13. Scitable, http://www.nature.com/scitable
14. Tobin, J. (2009). Semantic Social Media: The Future of Social Networking? A Special Podcast. March 2014.
15. Dandelion API, https://dandelion.eu/
16. Post object https://developers.facebook.com/docs/graph-api/reference/post

Semantic Publishing of Namespace Content and Trends in Search Engine Optimization in Petőfi Literary Museum

Anikó Mohay[1], Zsolt Bánki[1], and András Simon[2]

[1] Petőfi Literary Museum, Budapest Hungary
mohay.aniko@pim.hu; banki.zsolt@pim.hu
[2] Monguz Ltd., Szeged, Hungary
asimon@monguz.hu

Abstract: By appearing on the semantic web the artifacts and authors of the museum not only appear on the world wide web as autonomous entities but it also becomes possible to map their relations and connections. Through publicity, which in the future will not only concern our listing records of personal names, we increase the popularity and accessibility of our museum. There are new business models in the GLAM sector built on the initiative to publicize not only the metadata but our digital content too.

The main topic of this article focuses on the semantic web related namespace building activities of the Petőfi Literary Museum.

Keywords: namespace, authority data**,** semantic web, linked open data, open data, knowledge graph, knowledge vault, SEO, HTML5, Wikipedia, Wikidata, Petőfi Literary Museum.

1. A brief introduction to Petőfi Literary Museum

Petőfi Literary Museum (PLM) is the national museum for literary heritage in Hungary. It is one of the most respected museums in Hungary. The history of the institute goes back to the beginning of the 20th century; it collects the tangible and intangible heritage of Hungarian writers. The museum is open to new trends and methods in the contemporary cultural heritage field and accordingly, IT technology has been present in a wide range of our activities since the 1980's. We have been working on the development of our collection databases for decades, and now service the entire research environment. Our integrated collection management system meets all contemporary standard international requirements and we use it for describing our collection items and to develop significant namespaces. The most important part of this system is the personal name database, which contains approximately 600,000 authority records: it is the largest such dataset in Hungary. This namespace contains controlled and certified data stemming mostly from printed resources, and it is published through the online public access catalog (OPAC) of the museum (*opac.pim.hu*). The personal name space serves as the primary source for person name identification in the Hungarian library and museum system.

Our namespace includes different database genres (biography, bibliography, and genealogy). We developed several thematic databases for researchers which are unique in Hungary. Our colleagues working on the project seek to use only certified resources and have registered every difference among them, so the data are extremely useful for scholarly research and serve as a reliable reference.

Most of the databases based on lexicographical authority records are biographical, which presents information concerning the life and career of the person in question.

2. The Utilization of PLM Namespace – the Concept of the Hungarian National Namespace

In 2012 the experts of the Hungarian National Digital Archive, the National Széchenyi Library, the National Archive of Hungary and the Petőfi Literary Museum recognizing the professional utilization of the already existing significant namespace (Getty, VIAF, ICONCLASS) started to work out the concept of a National Namespace based on a collaborative principle. It was evident for them that the future is for the use of name authority files published on semantic web, which serves the demands of both public and archive users. The experts also agreed that the bases of the common authority file have to be provided by the individual institutional authority files and certain high quality archive name authority files could form the starting point for this. The project for creating the Hungarian National Namespace is in progress and the major part of the theoretical planning has been already completed. The theoretical plan of the National Namespace project states that the true meaning of creating the National Authority File is to make the cultural collection databases accessible through the collective archival use of authority files. We can only show the totality of Hungarian culture, all of our cultural values in their complexity and interconnection if we link our cultural collections through the operation of the common name authority files. Until the cultural data are stored in separate institutions as the public collection logic dictates we can only provide insular access for the public. The goal for the visitors is to be able to see and use all of the Hungarian cultural databases as a single integrated system instead of going from archive to archive to collect the data they are interested in. Considering this principle we started to work on the publication of the personal name authority records of PLM.

3. Open Personal Name Authority Records in PLM

3.1 Publication

The „developers" of the PLM namespace file can see the professional change by which the period of the isolatedly built institutional name authority files is replaced by the collaboratively built, for professional communities, freely useable data content. IT and the development of semantic web services have given the opportunity for the Petőfi Literary Museum to make its name authority file available and usable in the way presented below for the first time in Hungary.

3.2 Rules of using data

The Petőfi Literary Museum provides the audience with all the personal name authority data of the online catalogue applying the „CC0 1.0 Universal (CC0 1.0) Public Domain Declaration" license.

The museum uses the open data „badge". This determines, that our data can be freely used, recycled and redistributed by anyone - along with the source designation.

Based on the license PLM does not claim any copyright and waives its rights connected to the authority data. The museum does not take sanctions against additional users of the public data in concern. PLM does not take responsibility for any damage caused by the improper use of the public data.

3.2.1 Five stars of open data

The museum also uses „five stars of open data" badges founded by Tim Berners-Lee. The PLM now uses three of these which are: open licence, reuse, open format. The aim of the museum is to reach the maximum five stars that are already in the „link your data" concept.

Fig. 1. Record of Karinthy Frigyes (Hungarian writer) on the online catalog with open data badges (Source: *https://opac-nevter.pim.hu/en*)

3.3 Access

Authority data can be accessed through multiple channels, such as online catalog, OAI-PMH server, Z39.50 server. Records are published in widely used standard formats: Binary MARC (HUNMARC; MARC 21) MARC XML (HUNMARC; MARC 21) OAI DC XML, LIDO etc.

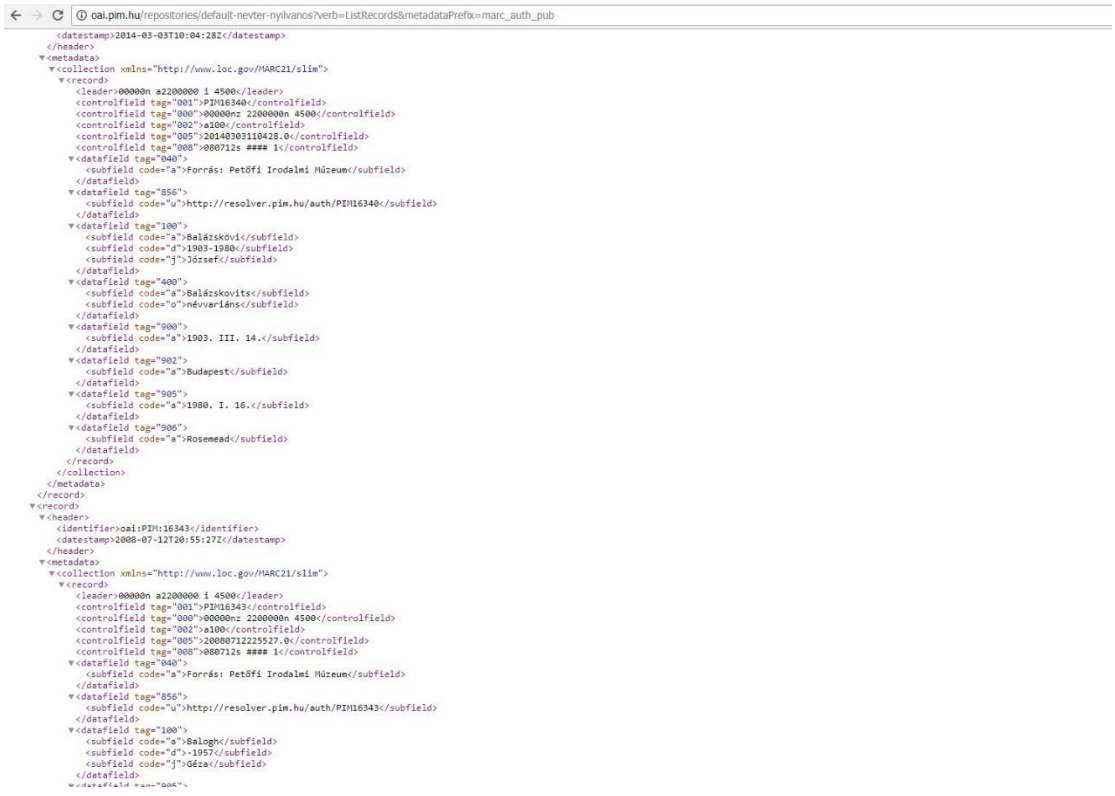


Fig. 2. Records from PLM OAI server (Source: http://oai.pim.hu/repositories/default-nevter-nyilvanos?verb=ListRecords&metadataPrefix=marc_auth_pub)

4. OPAC –Search Engine Optimization (SEO)

4.1 Search Console

We registered the URL of the OPAC in Google SearchConsole in order to follow the activities of indexing bots and to take the necessary SEO steps in the OPAC service based on the feedbacks.

4.2 Google Analytics- monitoring user activity

Google Analytics is the indicator of the effectiveness of online content streaming, which provides statistics about, number of users and page views, and the duration of their stay. These feedbacks will help with the search engine optimization.

5. Search Engine Optimization (SEO)

One way of appearance on the semantic web is the placement of semantic markers in the HTML page, which makes our site more interpretable for the search engines. With this method, we do not only make the "things" more understandable for the machine (which is the basic function of the semantic web) but we also support the search optimization.

In addition to inserting markers, further steps as placing formalized keywords in HTML, on Wikipedia, and on Wikidata will help our site to appear on the Web as part of the Knowledge Graph. The Knowledge Graph, which is described in more detail in the next chapter, is a knowledge network that creates relationships among the things in the world. We have implemented the steps of the search optimization and the appearance on the semantic web, with the technology partner of the Petőfi Literary Museum the Monguz Ltd., which is detailed in the following chapters.

5.1 The Concept of Search Engine Optimization-syntax optimization

The goal of search engine optimization (Search Engine Optimization= SEO) is that the search engines prefer certain websites as high as possible and move our content to the first page of the search results. To achieve this, we have the above-mentioned tools at hand, part of which are the HTML correction, structured data complemented by keywords, presence on Wikipedia and Wikidata.

By publicizing and making our name space, which was described in the previous chapters, widely available we increase/d the number of references by which we also helped our appearance in the Knowledge Graph and the knowledge panel (For some searches, you will see a Knowledge panel to the right of the page search results) in the search engine.

5.2 Knowledge Graph and Knowledge Vault – Semantic optimization

The Knowledge Graph and the Knowledge Vault are two new technical developments whose further building and enhancement the search engines continuously deal with.

Knowledge Vaults are typically large size databases in which the stored data use such semantic dictionaries as for example the schema.org dictionary. Their aim is to store millions of entities together with their connections.

Main Developers:
- Google Knowledge Graph
- Bing's Satori (Microsoft)
- Freebase (blended to Google Knowledge Graph)
- DBPedia
- Wikidata

As the most important developer of this semantic technology is Google and it is the most widely used one so the museum in its developments focuses on the Google Knowledge Graph and its successor the Knowledge Vault.

At the moment the Knowledge Vault is a service by Google. The Knowledge Graph is a general data web model that applies extended, large databases which store facts and information and the connections between them by using semantic dictionaries. The graph is searchable with the help of an easily available API. We get the results of the search back in a JSON-LD format where in the results we cannot find the connections between the entities as we could only get them from a full RDF graph.

5.3 HTML Correction

Searching for keywords has been present for years now. Search engines like Google indexes websites by the keywords of the site and then it ranks these among the search results. We integrated such keywords into certain attributes by which our users can search for our pages.

These keywords are present in:
- Title
- Heading1
- Content twice or 3 times
- Meta descriptions
- Picture names, alt text
- URL

We strived to use synonyms and compound keywords instead of accumulating keywords.

5.4 Semantic Elements in HTML5

Through structured data search engines or smart bots refer to the information located on our website and by this search engines show what entities our website contains.

We defined the different parts of our website by using the following semantic elements:

- <header> – Specifies a header for a document or section
- <main> – Specifies the main content of a document
- <nav> – Defines navigation links
- <section> – Defines a section in a document
- <summary>-Defines a visible heading for a <details> element
- <aside> – Defines content aside from the page content
- <figure> – Specifies self-contained content, like illustrations, diagrams, photos, code listings, etc.
- <mark> – Defines marked/highlighted text
- <figcaption> – Defines a caption for a <figure> element
- <footer> – Defines a footer for a document or section
- <time> – Defines a date/time

5.3 Schema.org

The Schema.org is a standard dictionary of „things" founded by Google, Yahoo and Yendex., are developed by an open community process, using the public-schemaorg@w3.org mailing list and through GitHub. In order that search engines understand what real life things our website describes (people, places, works) we used this dictionary.

The Schema markup is a collectively created marking system by which I can mark data in a structural way. For instance, I mark an entry as a person or an organization. The goal of this is to tell Google even more clearly what our data and content mean and what they are connected to.
Microdata a general term used to described the embedding of structured data within the HTML markup already on a page.

JSON-LD, structured data in web pages, usually in the header of the page.
RDFa, a specific part of RDF designed for embedding in HTML markup.

```
data-target="#accordion_relationships" aria-expanded="true"> <i class="fa
fa-angle-right" aria-hidden="true"></i> Rokoni kapcsolatok: </button>
      <div id="accordion_relationships" class="opac-accordion-wrapper collapse in"
aria-expanded="true" style="">
            <div class="opac-table record-accordion-table">
            <div class="opac-table-row accordion-row metadata-wrapper
accordion-wrapper-relationships">
                  <div class="opac-table-cell accordion-cell
metadata-name">Rokoni kapcsolatok: </div>
                  <div class="opac-table-cell accordion-cell metadata-value"><div
property="relatedTo" type="Person">
      <span property="name"class="textComponent reference"
onmouseover="getReferenceAttributes(this);" referencetype="bibl_record" host=""
idname="BIBID-CONFIGURABLE" idvalue="PIM470729">Pásztor Mária 1858-1937: anya
      </span>
</div><br><span <div property="relatedTo" type="Person"><span
property="name"class="textComponent reference"
onmouseover="getReferenceAttributes(this);" referencetype="bibl_record" host=""
idname="BIBID-CONFIGURABLE" idvalue="PIM1113432">Ady Lőrincné
1858-1937</span></div><br><span <div property="relatedTo" type="Person"><span
property="name" class="textComponent reference"
onmouseover="getReferenceAttributes(this);" referencetype="bibl_record" host=""
idname="BIBID-CONFIGURABLE" idvalue="PIM235648">Ady Lőrinc 1851-1929:
apa</span><br></div><span <div property="relatedTo" type="Person">
<span property="name"class="textComponent reference"
onmouseover="getReferenceAttributes(this);" referencetype="bibl_record" host=""
idname="BIBID-CONFIGURABLE" idvalue="PIM48200">Boncza Berta 1894-1934</span></div>
            </div>
            </div>
      </div>
```

Fig. 4. RDFa elements in our HTML

5.4 Wikipedia - Wikidata

Wikipedia and Wikidata are two of the most important elements of SEO, from which the Knowledge Graph builds up too. In these databases PLM is registered and at certain personal names in the entries PLM related OPAC links appear. Our museum has already contacted the developers of Wikipedia and Wikidata who are showing the suitable OPAC links in Wikipedia in personal entries after identifying it based on name, place and time of birth and death. Then they imported the references created in this way to Wikidata.

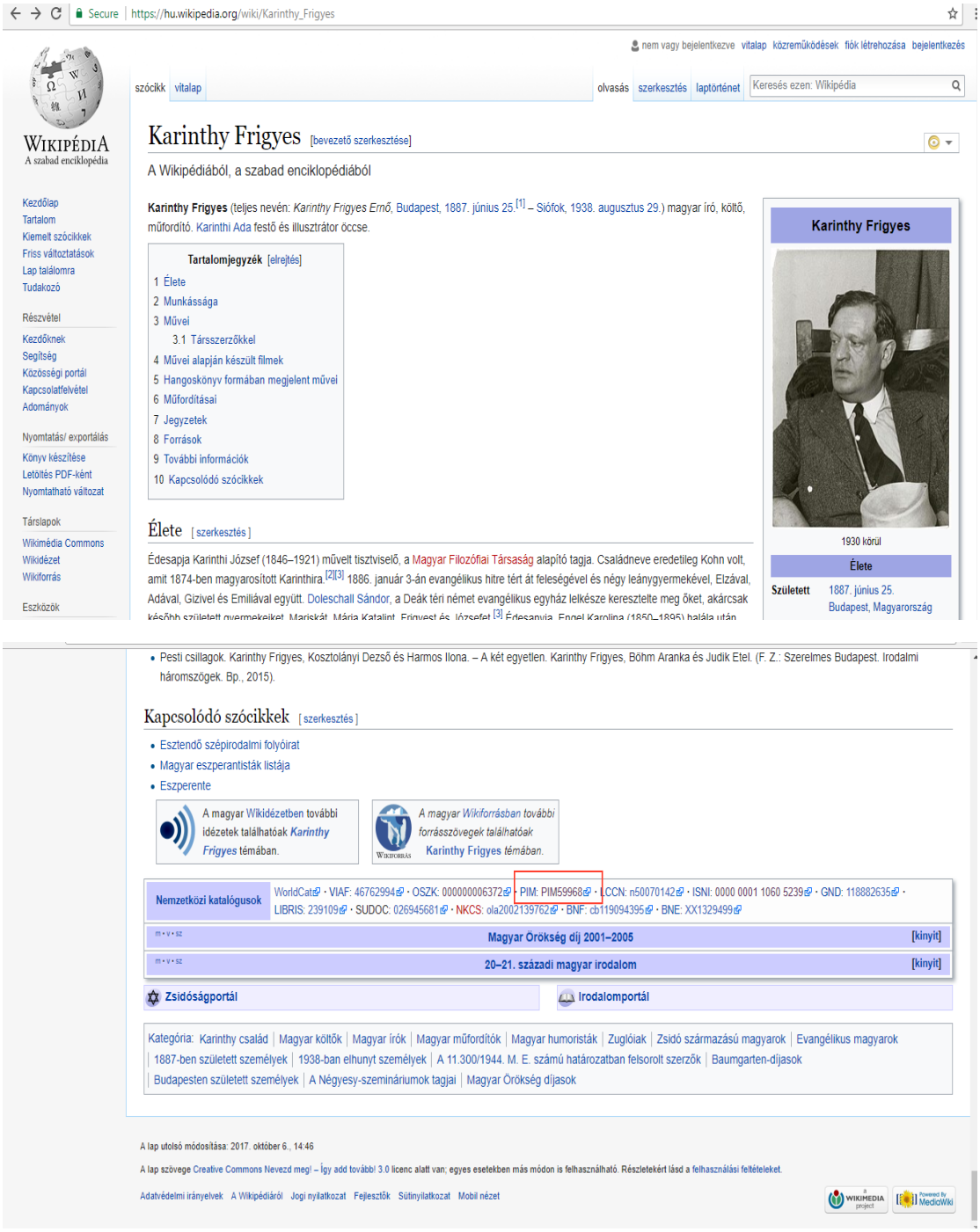


Fig. 5. PLM reference in the Wikipedia

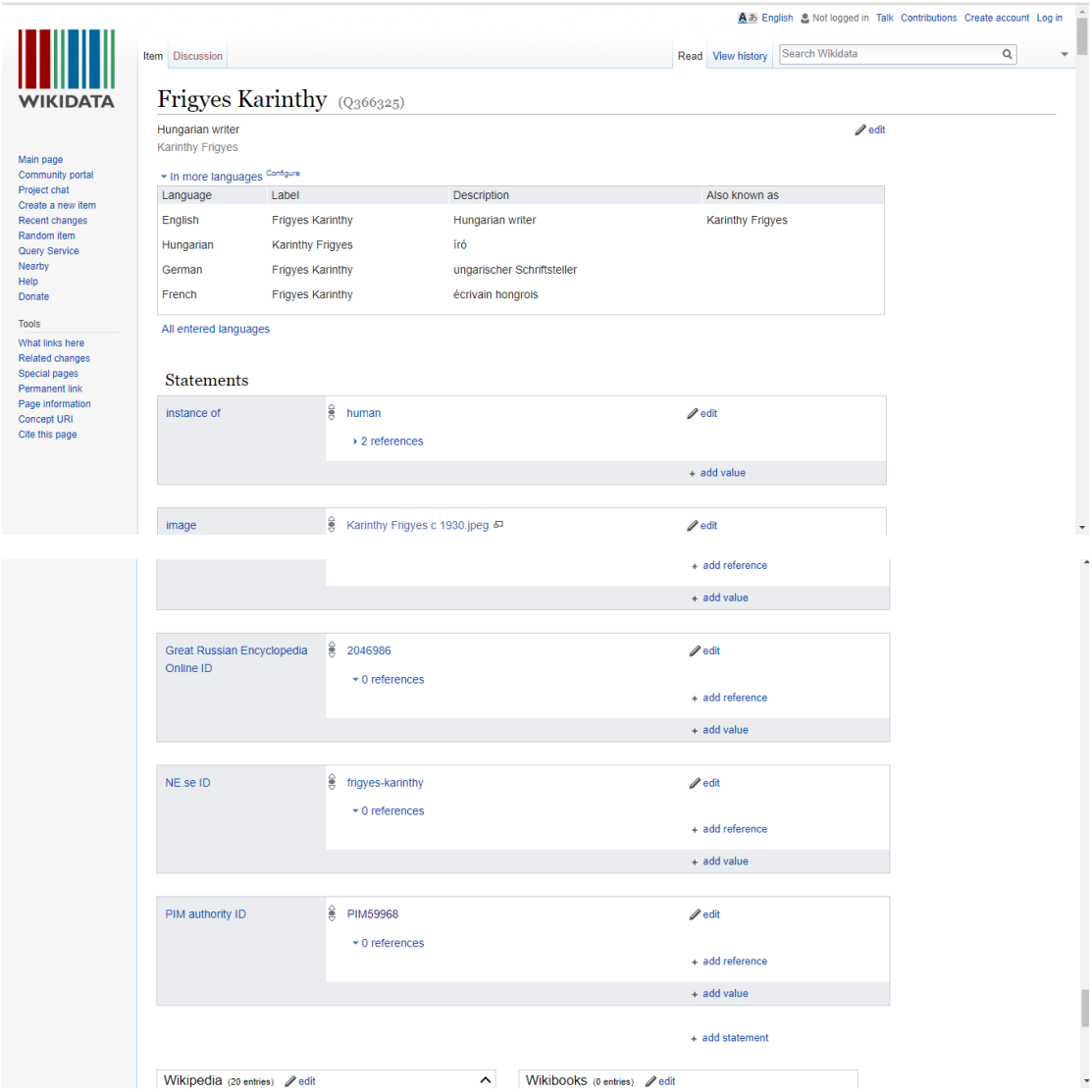


Fig. 6. PLM in the Wikidata

However, it is important to create the semantic background (in progress). At the moment the developers are working on a way to automatically match the personal name entries with the OPAC links of the museum and then move them to Wikipedia. It is very important to increase the number of references, there is an index for this called PageRank (it is not public any more but it is used by Google as an algorithm). PageRank is a numeric indicator of importance based on the number of links pointing to a particular webpage. So, every link from other websites pointing to any page of our website increases the value of the PageRank index. Wikipedia links got into the PLM database in the proper MARC field.

The indicator of schema.org is placed in the HTML page by which we point to the Wikipedia equivalent of a particular personal name

The next step of developments is to create the RDF XMLs and the formation of graphic visualization so that the users can discover the new and existing connections.

References

1. Bánki, Zs., Lengyel, M., & Tóth, K. (2007). “Ablak” a múzeumokra: A Petőfi Irodalmi Múzeum speciális adattárai a Huntékában. Networkshop 2007. Eger: NIIF. from https://nws.niif.hu/ncd2007/docs/aen/095.pdf

2. Bánki, Zs., Mészáros, T., Németh, M., & Simon, A.: Azonos személynevekre vonatkozó besorolási rekordok automatikus felderítése a PIM adatbázisában (2016). In: TMT 2016 (63.évf) 12 sz. pp.471-477

3. Creative Commons https://creativecommons.org/publicdomain/zero/1.0/deed.hu

4. DataHub. https://datahub.io/about

5. Horváth, Á. (2011). Linked Data at NSZL. http://nektar.oszk.hu/w/images/0/04/LinkedDataAtNszl_06.pdf

6. Jason A. Clark and Scott W. H. Young (2016): Linked Data is People: Building a Knowledge Graph to Reshape the Library Staff Directory In: Code4Lib Issue 36, 2017-04-20 http://journal.code4lib.org/articles/12320

7. Linkeddatatools http://www.linkeddatatools.com

8. Ronallo, Jason. “HTML5 Microdata and Schema. org.” Code4Lib Journal 16 (2012). http://journal.code4lib.org/articles/6400

9. Schema.org. http://schema.org/

10. 5 star open data http://5stardata.info/en/

11. Creative Commons https://creativecommons.org/publicdomain/zero/1.0/deed.hu

12. w3school.com. https://www.w3schools.com

A Study on the Construction Technology Digital Library Service Information Model: focusing on the case of Korea's Construction Technology Information System

Seong-Yun Jeong

Korea Institute of Civil Engineering and Building Technology, Research Institute for ICT Convergence & Integration, Gyeonggi-Do, South Korea
syjeong@kict.re.kr

Abstract. To develop a customized information service and information search function for application in South Korea's Construction Technology Information System, this study presented an information model for metadata for construction technology digital library, construction business, WBS, and user-customized services.

Keywords: metadata for construction technology digital library, XML Schema, construction project information, work breakdown structure, user-customized service, information models.

1 Introduction

The South Korean government enacted the Construction Technology Promotion Act (CTPA) in 1988 to promote technology development in the construction and engineering industry. Article 18 of CTPA contains a comprehensive distribution system of construction technology information for sharing and utilizing the results and experiences of construction technology development so that small and medium-sized construction and engineering companies can strengthen their technological competitiveness. In accordance with Article 18 of CTPA, the Korea Institute of Civil Engineering and Building Technology (KICT) developed the Construction Technology Information System (CTIS) in 2001 and has since been operating the system to collect construction technology related digital documents, transform them into a database, and allow the users to access the database.

As the number of users of CTIS increases and as the number of construction technology related digital documents rises, the demand for the easier finding of the desired information and for obtaining more accurate information has begun to grow. This study developed a user-customized service function to satisfy such need of the users. To efficiently manage metadata as a service function development, this study proposed the Metadata for Construction Technology Digital Libraries (M4CODIL), which integrate the construction report metadata and the metadata for construction project working data being used by CTIS.

In addition, to provide the users with ample bibliographic information on construction technology related digital documents, this study defined the element sets that constitute information models as well as information models for construction projects, WBS, and user-customized services.

Further, to examine the applicability of the information models proposed herein, bibliographic information on the data of construction cost reduction cases was created as an XML instance according to the XML-Schema-based information model. Finally, the limitations of this study and a need for further research were pointed out.

2 Analysis of Related Cases

2.1 Construction Technology Information System

CTIS is a construction technology information portal system developed for searching construction technology related digital documents in one place. CTIS provides a database of bibliographic information on about 45,000 construction technology related digital documents, including construction project reports, construction technology development reports, cases of construction technology applications, cases of construction project cost savings, and standardized estimating unit manpower and materials. The database is open to anyone who wants to search such bibliographic information [1].

2.2 Metadata of Construction Reports and Construction Project Working Data

Metadata are data about other data, and they generally also refer to the bibliographic information (lists, indices, etc.) on books being used in conventional libraries. With the wide spread of the Internet, digitalized data are increasingly being created in large numbers, and as such, for searching and managing the necessary digitalized data, metadata standardization is essential.

To accurately express the bibliographic information, KICT is using the basic element set defined in Dublin Core's Metadata Element Set was used [2], and considering the characteristics of construction technology related digital documents, element sets like "ProjectOrganization," "MetaMetadata," "KC Subject," "Collection," and "Open" were added to it [3].

In addition, of the encoding schemes used in element sets, "KICTWC" and "KICTGC" are the abbreviations of KICT-Web Classification and KICT-Green Classification, respectively, and they refer to the classification systems that indicate the characteristics of the relevant data defined by KICT.

"KICT-Contributor" refers to the institutional contributor code classified by KICT, and "KICT-Open" refers to the classification code that defines the level of opening of relevant data. "KICT-Type" is the data type classification code defined by KICT. "KCICF," "KCICS," "KCICE," and "KCICW" are the construction information classification systems announced by South Korea's Ministry of Land, Infrastructure, and Transport (MOLIT), and they refer to five categories, such as facility classification (F), space classification (S), Element classification (E), work type classification (W), and resource classification (R).

2.3 Work Breakdown Structure Information System

MOLIT started to develop WBS in 2008 to manage the progress of the road and river construction projects that it ordered itself. WBS defines the scope of work necessary for systematically managing the progress of construction projects, and identifies the individual work elements contained in the scope of work. It refers to the criteria by which, among the identified work elements, the elements with similar characteristics are integrated into one category and grouped and classified hierarchically. As shown in Table 1, WBS has a top-down hierarchical structure in which the elements of work are classified in greater detail from the upper level (level 1) to the lower level (level 6) [4].

The WBS used in the construction of roads and rivers is classified into six stages, and it collects the elements of work to be carried out at each classification stage in the relevant construction project. As such, there is a close relationship between the construction project information and the WBS, and the performance process and results of each work component are used to create the drawing and document outputs for construction execution.

Table 1. An example of WBS for a road construction project

Level	WBS	Task Example
1	Facility Class	road, structure, other facilities, etc.
2	Major Works Class	earth, drainage, pavement, secondary, etc.
3	Facility Class by Works	main line, branch line, IC, underpass, etc.
4	Directional Area Class	upward, downward, common, etc.
5	Expanded Area Class	section name, road name, lamp name, etc.
6	Work Package	excavation, closed conduit, transportation facilities, etc.

2.4 Element Set of Construction Projects

This study sought to link the construction technology digital library metadata and the construction project information so that the users can search construction technology related digital documents more easily and accurately.

For this purpose, among MOLIT's systems and guidelines, the construction project management system, construction project post-evaluation system, guidelines for the digital delivery of drawings and documents for construction execution, and guidelines for quantity calculation classification system that are directly related to the road and river construction projects were selected [5]. The element sets related to the construction projects used in these systems and guidelines were examined and compared, as shown in Table 2.

Table 2. An example of WBS for a road construction project

Category	Element set
Construction Project Management System	Site name, site classification, project name, project code, contractor information (business registration number, company name), site location (address), contract information (agreement date, construction commencement date, projected completion date, bid date, bid rate, successful bidder’s price), site manager (name, login ID, contact point), etc.
Construction Project Post-Evaluation System	project name, project type, project size, characteristics of construction, area of construction, client information, project outline, subcontracting method, nature of contract, bid method, contracting method, contractor name, etc.

Guidelines for the Digital Delivery of Drawings and Documents for Construction Execution,	project name, construction name, contact name, project type, construction stage, construction commencement date, projected completion date, client information(organization name, staff name, phone number), contractor information (contractor name, staff name, phone number), class of security, project management number, etc.
Guidelines for the Quantity Calculation Classification System	project name, construction commencement date, projected completion (planning) date, bid date, bid method, subcontracting method, successful bidder's price, bid rate, contracting method, nature of contract, etc.

3 Information Models for Construction Technology Digital Library Services

3.1 Establish Relationships between Information Models

Before presenting the information model for the construction technology digital library service, this study established the following preconditions for the effective design of the information model.

First, construction technology digital library metadata were constructed using basic element sets to efficiently manage the bibliographic information of the construction technology digital library. Of course, the more the element sets, the more the bibliographic information of the construction technology digital library can be fully expressed. This, however, will increase the amount of work required to collect and process data and to construct a database that corresponds to the element sets, as well as the efforts needed to manage the data accordingly and the cost of managing such data. Therefore, only the minimum element sets needed to retrieve the data were included in the construction technology digital library metadata.

Second, the South Korean government strictly controls the collection of personal information in accordance with CTPA. To comply with this government policy, element sets were selected focusing on the service information that the users need rather than on the personal information of the users.
Third, as it was assumed that the results of this study would be applied to CTIS, the information systems and element sets that are currently being used were used as they are. The element sets that were redundant or ineffective were removed, however, and element sets to be used to ensure the connectivity between information models were added.

This study redefined the construction technology digital library metadata that combine each of the metadata for construction reports and construction project working data to provide the users with better convenience under these preconditions. The information models needed to provide convenience were also added, and the information systems and element sets in the information models were defined. In this process, element sets were added for the relational keys used to link the construction technology digital library metadata and information models.

Fig. 1 shows the linking relationships between information models like construction projects, WBS, and CTIS members, focusing on construction technology digital library metadata. The attributes of the information models were defined as having a one-to-many (M) relationship, taking into account the correlations, such as concerning how many times the data are referenced between the information models.

Again, element sets were established for the relational keys used between the construction projects and the WBS information models and between the CTIS members and the user-customized service information models, and the attributes of these information models were defined as having a one-to-M relationship.

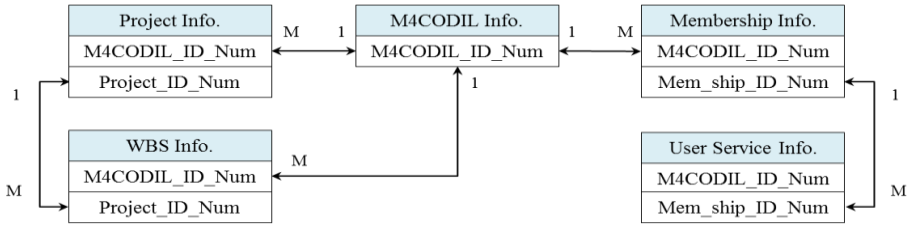


Fig. 1. Correlation between information models

3.2 Information Model for Metadata for the Construction Technology Digital Library

The construction report metadata and the construction project working data metadata basically use the metadata element set defined in the Dublin Core [2]. The metadata element set is a set of bibliographic information that defines the basic elements (e.g., title, author, publisher, data format, language, etc.) needed to efficiently search and manage digital resources.

This study determined that it is desirable to redefine the metadata elements as M4CODIL in the future considering the compatibility with the information services provided by other similar systems. In addition, among the element sets constituting these metadata, there exists a plurality of element sets used with the same or similar meaning among the element sets defined considering the nature of the construction technology related digital documents. So, it was determined that it is efficient to integrate and manage metadata rather than to manage individual metadata.

The “Project Organization” element has the same meaning as the element set defined in the construction project information model. As the “KC Subject” element can be used to mean the “Facility Class” and “Major Works Class” elements of the WBS, these element sets were excluded from the construction technology digital library metadata.

Fig. 2 shows the XML schema structure of the information model of the construction technology digital library metadata.

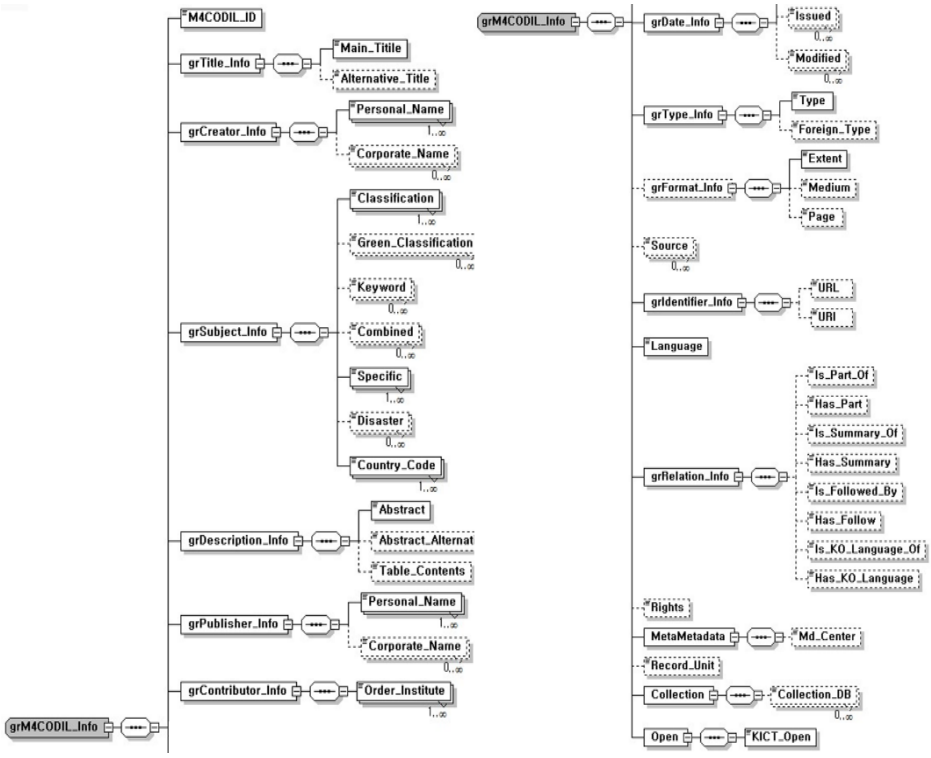


Fig. 2. Example of XML schema structure for construction technology digital library metadata

3.3 Information Model for Construction Projects and WBS

This study extracted only the element sets that can be used to search the construction technology related digital documents or to determine which construction technology related digital documents are associated with a construction project based on the element sets of the construction project.

Meanwhile, an information model consisting of an information system and element sets was designed for WBS for accurately searching construction project working data like process procedures and cases of construction project cost saving. The WBS information model defined an element set that is used as a relationship key so that it can be directly linked to the construction technology digital library metadata in case of the absence of a construction project, such as standards for construction projects. Conversely, if there is a corresponding construction project working data, such as cases of construction project cost savings, an information model was defined to link the construction technology digital library metadata via the information model of the construction project [6].

Fig. 3 is information models that express the element sets of WBS, and their attributes in XML Schema.

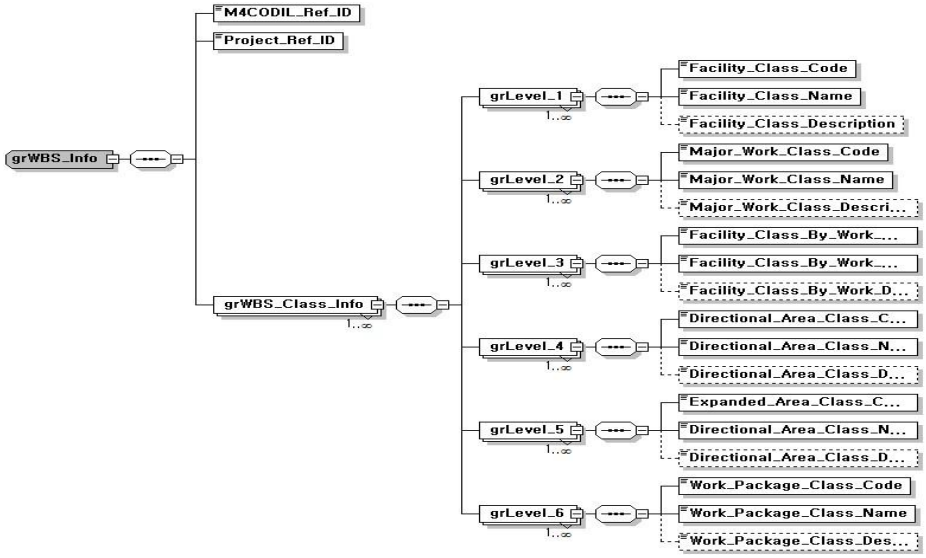


Fig. 3. Example of XML schema structure for WBS

3.4 Information Model for User-customized Services

To provide a user-customized service, it is basically necessary to secure the personal information about the user who receives the service. Collecting personal information, however, is strictly regulated according to CTPA, so it was deemed desirable, instead of indicating the personal information of the user and the user-customized service information together in the information model for the CTIS members, to collect the information that is useful for the user, apart from the CTIS member information, or to automatically create the information from the database and show it.

Considering these limitations, this study decided to use the element sets of members that are currently being used in CTIS as they are, while separately defining the information model for the user-customized service. In addition, element sets for the relational keys were added to link the information models of CTIS members and the user-customized service.

This study defined the element sets for user-customized services as shown in Table 3 and Fig. 4 by collecting opinions through experts' advisory meetings in October 2016, as well as by surveying 312 CTIS users in June 2016.

Table 3. Element set of user-customized services

Element Set		**Date Type**	**Encoding Scheme**	**Necessary**	**Repetition**	**Remarks**
Element Set	**Date Type**					
M4CODIL identification number		string		○	×	Automatic creation of DB
CTIS use information	Total hits	integer		○	×	
	Total access to original texts	integer		○	×	
	Total searches	integer		○	×	
Popular		string		○	×	

search words	Recommended areas	string	KICTWC	○	○	
Areas of interest related to technology		string		○	○	Input directly
	Latest search words	string		○	○	Automatic creation of DB
	Popular search words among other users			○	○	

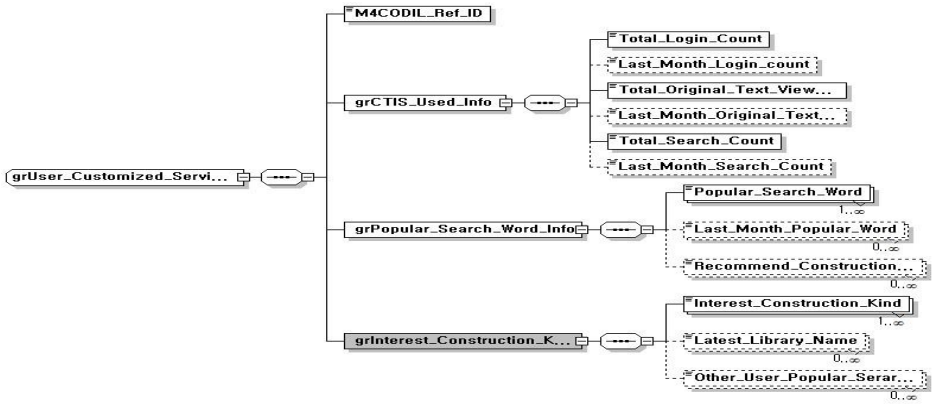


Fig. 4. Example of XML schema structure for user-customized services

The attributes of the element set of “areas of interest related to technology” in Table 3 were defined to allow new members to select one or more areas of interest from the list of categories of technology that determine the category of members. The attributes of the other element sets were designed to allow the automatic extraction of the necessary data from the log database of CTIS.

3.5 Review of Applicability of Information Models

To The proposed designed information models were reviewed based on the cases of cost saving provided by CTIS, using XML. XML can accurately express the attributes of construction technology digital library metadata, and in a separate XML conversion process, it can automatically store the data recorded in instances in the database, or can conversely transform the data stored in the database into instances.

Due to such advantages, XML was selected as the electronic document exchange standard format for the construction project informatization project now under way in South Korea.

The afore-defined element sets and their attributes constituting the information models were developed in XML Schema using the XML syntax rules. Data corresponding to such element sets were created according to the developed XML Schema, and were created into instances. In this process, it was reviewed if the element sets and their attributes could be properly reflected in XML Schema.

In addition, it was confirmed if the bibliographic information on the data of cases of construction project cost saving could be properly transformed into instances in accordance with XML Schema.
Fig. 5 shows an example of an XML instance that shows linkages among M4CODIL, the construction project information model, and the WBS information model.

```
<WorkBreakdownStructureInformation xmlns="http://www.codil.or.kr/XMLPool/RootSchemaMod-
ule" :wbs="http://www.codil.or.kr/XMLPool/MetadataforConstructionTechnologyDogotalLibraryEntitiesSchema-
Module" :xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLoca-
tion="http://www.codi.or.kr/XMLPool/M4CODIL_Info.xsd">
 <!-- XML Instance for Work Breakdown Structure Information-->
  <grWBS_Info>
    <!-- Reference ID of Construction Technology Digital Library -->
  <wbs:m4coddil_ref_ID type="xsd:string minOccurs="1" maxOccurs="1">PRA-2015-0355
  </wbs:m4coddil_ref_ID>
     <!-- Reference ID of Construction Project -->
    <wbs:project_ref_ID type="xsd:string minOccurs="1" maxOccurs="1">SEL-2013-0342
    </wbs:project_ref_ID>
    <wbs:grWBS_Class_Info>
     <wbs:grLevel_1 refine="facility_class_code" type="xsd:string" minOccurs="1"
      maxOccurs="1"scheme="KICTWBS">F11120</wbs:grLevel_1>
       <wbs:grLevel_1 refine="facility_class_name" type="xsd:string minOccurs="1"
      maxOccurs="1">Main Line</wbs:grLevel_1>
       <wbs:grLevel_1 refine="facility_class_description" type="xsd:string" minOccurs="0"
      maxOccurs="1">Technological development of median strip with high quality
     </wbs:grLevel_1>
          <! --The remaining elements are omitted.--->
```

Fig. 5. Example of expressing the linkage between information models as an XML instance

4 Conclusion

With the ever-increasing numbers of CTIS users and construction technology digital libraries, there are increasing requirements for the search function for faster data search and information that is useful for every user.

To meet such user requirements, this study developed a user-customized service function. Towards this end, integrated metadata combining the construction reports metadata and the construction project working metadata were presented. In addition, information models used to inquire about information on construction technology digital-libraries-related construction projects and WBS, and to provide useful information services to the users, were newly defined. The defined information models were developed in XML Schema, and their applicability was confirmed.

The proposed XML-Schema-based M4CODIL was judged to be more effective than the existing information services in terms of user-customized services and the information search function.

The proposed information models, however, are still at the pilot stage prior to the launch of construction technology digital library services through CTIS. Thus, before launching the full-swing information services, there is a need to construct detailed information models and information systems that will fully reflect the features of construction technology related digital documents, as well as to conduct further research for accurately expressing the meanings of the attributes of element sets.

Additionally, there is a need to develop a program for the automatic conversion between DB and XML, and to redesign the CTIS DB was based on the proposed information models.

References

1. Construction Technology Information System, http://www.codil.or.kr (Written in Korean)
2. Dublin Core, http://dublincore.org/metadata-basics/
3. An, S., Cho, J.G., Kim, S.Y.: 14 Operation of the Construction Technology Knowledge Information Database and Service System. Research report, pp. 64-65, KICT press, South Korea (2014)
4. Park, H.P., et al.: A Study on the Introduction of a Code System for Construction Standards. Research report, pp. 139-173, KICT press, South Korea (2013)

5. Korea Construction Project Informatization (Continuous Acquisition & Life Cycle Support), https://www.calspia.go.kr/data/selectOperationReportList.do (Written in Korean)

6. Jeong, S.Y.: A Study on the Introduction of the Work Breakdown Structure for Infrastructure Asset Management. In: 6th International Conference on Construction Engineering and Project Management, pp. 691-692, South Korea (2014)

Author Index