

Challenges and Opportunities for Knowledge Organization in the Digital Age

Proceedings of the
Fifteenth International ISKO Conference
9-11 July 2018
Porto, Portugal

Organized by
International Society for Knowledge Organization (ISKO),
ISKO Spain and Portugal Chapter
University of Porto – Faculty of Arts and Humanities
Research Centre in Communication, Information
and Digital Culture (CIC.digital) – Porto

Edited by
Fernanda Ribeiro
Maria Elisa Cerveira

Advances in Knowledge Organization, Vol. 16 (2018)

Challenges and Opportunities
for Knowledge Organization
in the Digital Age

Proceedings
of the
Fifteenth International ISKO Conference
9-11 July 2018
Porto, Portugal

Organized by
International Society for Knowledge Organization (ISKO),
ISKO Spain and Portugal Chapter
University of Porto – Faculty of Arts and Humanities
Research Centre in Communication, Information
and Digital Culture (CIC.digital) – Porto

Edited by

Fernanda Ribeiro
Maria Elisa Cerveira

ERGON VERLAG

Editorial Support:

Raquel Graça

Predocumentation:

The volume contains: Introduction – Keynote Address – Foundations and Methods for Knowledge Organization – Interoperability towards Information Access – Societal Challenges in Knowledge Organization – Poster – Workshop – List of Contributors and Authors' Index

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche
Nationalbibliografie; detailed bibliographic data are available in the
Internet at <http://dnb.d-nb.de>.

© Ergon – ein Verlag in der Nomos Verlagsgesellschaft, Baden-Baden 2018

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in other ways and storage in databanks.

Duplication of this publication or parts thereof is only permitted under the provisions of the German Copyright Law, a copyright fee must always be paid.

Cover Design: Jan von Hugo

www.ergon-verlag.de

ISBN 978-3-95650-420-4 (Print)

ISBN 978-3-95650-421-1 (ePDF)

ISSN 0938-5495

Table of Contents

Introduction

<i>Fernanda Ribeiro</i> -----	15
-------------------------------	----

Keynote Address

Supporting truth and promoting understanding: Knowledge Organization and the curation of the infosphere <i>David Bawden</i> -----	17
--	----

Foundations and Methods for Knowledge Organization

Among dictionaries, reference works and concept systems: can the terminology in International Standard ISO 5127:2017 contribute to Knowledge Organization? <i>Axel Ermert</i> -----	29
--	----

An Analysis of the theoretical and practical application of Diplomatics to archival description in Knowledge Organization <i>Natalia Bolfarini Tognoli, Ana Célia Rodrigues</i> -----	43
--	----

Applied Knowledge Organization and the history of the world <i>Rick Szostak</i> -----	52
--	----

Approaches to the concepts of exhaustivity and specificity in ISKO International meeting proceedings: 2000-2017 <i>Maria da Graça Simões, Daniel Martínez-Ávila, Blanca Rodríguez-Bravo, Patricia de Almeida, Isadora Victorino Evangelista</i> -----	58
--	----

Articulating a cultural research program for Knowledge Organization Systems <i>Gregory H. Leazer, Robert Montoya, Jonathan Furner</i> -----	66
--	----

Aspects regarding the notion of subject in the context of different theoretical trends: teaching approaches in Brazil <i>Brisa Pozzi de Sousa, Cristina Dotta Ortega</i> -----	74
---	----

Automatic indexing and ontologies: the consistency of research chronology and authoring in the context of Information Science <i>Maria da Graça Simões, Luís Miguel Machado, Renato Rocha Souza, Maurício Barcellos Almeida, António Tavares Lopes</i> -----	86
---	----

Big data, Knowledge Organization and decision making: opportunities and limits <i>Amos David, Nadine Ndjock</i> -----	95
Classification and Knowledge Organization Systems: ontologies and archival classification <i>Thiago Henrique Bragato Barros, Daniel Libonati Gomes</i> -----	103
Classification of photographs: methodological concepts in Archival Science, Library Science and Museum Science <i>Ana Cristina de Albuquerque, Luciane de Fátima Beckman Cavalcante</i> -----	112
Comparing the use of full text search between a conventional IR System and a DBMS <i>Edson Marchetti da Silva</i> -----	120
Computer-assisted checking of conceptual relationships in a large thesaurus <i>Decio Wey Berti, Jr., Gercina Lima, Benildes Maculan, Dagobert Soergel</i> -----	128
The Contribution of Semiotics to Knowledge Organization for music information <i>Camila Monteiro de Barros, Lígia Maria Arruda Café, Audrey Laplante</i> -----	137
Critical questions for big data approach in knowledge representation and organization <i>Lala Hajibayova, Athena Salaba</i> -----	144
Developing a field of knowledge through bibliography: art history in the 16 th century <i>Giulia Crippa</i> -----	152
Different parameters for Knowledge Organization in archives <i>Sonia Troitiño</i> -----	160
Document representation with images: an experimental milestone <i>Marcilio de Brito, Widad Mustafa El Hadi, Maja Žumer, Simone Bastos Vieira, Marcos de Brito</i> -----	167
Domain analysis of scientific production in Information and Communication Technology in the context of small farmers <i>Fábio Mosso Moreira, Ely Francina Tannuri de Oliveira, Ricardo César Gonçalves Sant'Ana</i> -----	176
Epistemic communities, domain analysis, and Kuhn: dialogs and intersections in Knowledge Organization <i>Daniel Martínez-Ávila, Isadora Victorino Evangelista, Maria da Graça Simões, José Augusto Chaves Guimarães</i> -----	184
Epistemic <i>LOCI</i> : linguistic and critical meta-methodology in Knowledge Organization <i>Tatiana de Almeida, Gustavo Silva Saldanha</i> -----	191
Epistemological challenges in Knowledge Organization in the digital age <i>Rosa San Segundo, Maria Adelina Codina-Canet</i> -----	198

Facet: itself a multifaceted concept <i>Michèle Hudon, Alexandre Fortier</i> -----	204
Feasibility of implementing PRESSoo model in organizing Persian serials <i>Negin Shokrzadeh Hashtroudi, Mohsen Haji Zeinolabedini</i> -----	212
Foundations and methods for Knowledge Organization in European iSchools courses <i>Olivia Pestana</i> -----	224
A Framework to represent variables and values in Social Science research data sets to support data curation and reuse <i>Guangyuan Sun, Christopher S. G. Khoo</i> -----	231
The Genesis of Documentation in Brazil: Manoel Cícero Peregrino da Silva and Paul Otlet (1900-1924) <i>Carlos Henrique Juvêncio, Georgete Medleg Rodrigues</i> -----	240
Global Knowledge Organization, “super-facets” and music: universal music classification in the digital age <i>Deborah Lee, Lyn Robinson, David Bawden</i> -----	248
Historical ambiguity: a lens for approaching outdated terms <i>Lilium Rajan</i> -----	256
How Knowledge Organization helped to shape the emerging field of Terminology in Canada <i>Lynne Bowker</i> -----	265
Influence of metatheoretical research in Knowledge Organization <i>Paula Carina de Araújo, Joseph T. Tennis</i> -----	273
Information and argument structures in Sociology research abstracts <i>Wei-Ning Cheng, Christopher S. G. Khoo</i> -----	282
Information and knowledge ecology: a field for research in Knowledge Organization <i>Wiesław Babik</i> -----	290
Information from jussive processes <i>Elliott Hauser</i> -----	300
Intellectual history, history of ideas, and subject ontogeny <i>Joseph T. Tennis</i> -----	308
Is the massive incorporation of e-books into university libraries devaluing the technical processes related to the assigning of subject headings and classification codes? <i>Isidoro Gil-Leiva, Mariângela Spotti Lopes Fujita, Pedro Manuel Díaz Ortuño, Daniela Majorie dos Reis</i> -----	314

Knowledge Organization in editorial policies for titles, abstracts and keywords in JCR-indexed journals: an exploratory study in the areas of Information and Communication Sciences <i>Mariângela Spotti Lopes Fujita, María-del-Carmen Agustín-Lacruz, Ana Lúcia Terra</i> -----	321
Knowledge Organization in the health field: an ontology project to improve the information retrieval process <i>Claudio Jose Silva Ribeiro, Diones Ramos da Silva</i> -----	330
Literary warrant-based approach to organize KO terminology: criteria and method <i>Mario Barité, Mirtha Rauch, Ana Inés Brozia, Micaela Morales</i> -----	339
Medical ontology: Siddha System of Medicine <i>Hemalata Iyer, K.S. Raghavan</i> -----	347
The Modern and its impact on models of information organization in Brazil: the decline of the National Library and the rise of the National Book Institute (1930-1954) <i>Carlos Henrique Juvêncio, Georgete Medleg Rodrigues</i> -----	356
The Onomasiological approach and the function of definitions in the elaboration of domain models in ontologies <i>Maria Luiza de Almeida Campos, Hagar Espanha Gomes</i> -----	363
OntoM4IS: ontology reuse method for Information Science <i>Helder Noel Monteiro Firmino, Gercina Ângela de Lima</i> -----	371
Parsimony in Biological and Colon Classifications <i>Robert D. Montoya</i> -----	377
The Penumbra-line: Ranganathan's journeys and the genesis of the APUPA pattern <i>Luca Giusti</i> -----	380
Photography as a legitimate technique for domain analysis in Knowledge Organization <i>Eva Hourihan Jansen</i> -----	392
Reconstructionism: a comparative method for viewpoint analysis and indexing using the example of Kohlberg's moral stages <i>Michael Kleineberg</i> -----	400
A Relation typology in Knowledge Organization Systems: case studies in the research data management domain <i>Jian Qin</i> -----	409
Reorganization of knowledge in technical documents for the information system: the extraction of logico-cognitive information from heterogeneous administrative data <i>Omar Larouk, Mabrouka El Hachani</i> -----	416
The Retrieval power added by subject indexing to bibliographic databases <i>Philip Hider</i> -----	426

Risks and requirements of an applied Information Science research framework in healthcare <i>Fernanda Gonçalves, Gabriel David</i> -----	432
Semantic warrant, cultural hospitality and knowledge representation in multicultural contexts: experiments with the use of the EuroVoc and UNBIS thesauri <i>Roger de Miranda Guedes, Maria Aparecida Moura</i> -----	442
Solid Foundations and some secondary assumptions in the design of bibliographic metadata: toward a typology of complementary uses of metadata <i>Athena Salaba, Joseph T. Tennis</i> -----	450
The Spirit of inquiry's power to influence in 21st-century KO research: Jesse Shera and Margaret Egan <i>José Augusto Chaves Guimarães, Maria Claudia Cabrini Gracio, Daniel Martínez-Ávila, Rodrigo de Sales</i> -----	460
Truth, relevance, and justice: towards a veritistic turn for KO <i>Jonathan Furner</i> -----	468
Verbal protocols in Brazilian Information Science: a perspective from indexing studies <i>Paula Regina Dal' Evedove, Roberta Cristina Dal' Evedove Tartarotti, Mariângela Spotti Lopes Fujita</i> -----	475
Who is Tesouro? The man, words and things <i>Gustavo Silva Saldanha, Naira Christofolletti Silveira, Giulia Crippa and Tatiana de Almeida</i> -----	483
Wikipedia categories in research: towards a qualitative review of uses and applications <i>Jesús Tramullas, Ana I. Sánchez-Casabón, Piedad Garrido-Picazo</i> -----	490

Interoperability towards Information Access

A Comparative analysis and evaluation of bibliographic ontologies <i>Maria Teresa Biagetti</i> -----	501
A Conceptual model for an OWL ontology to represent the knowledge of transmedia storytelling <i>Juan Antonio Pastor Sánchez, Tomás Saorín Pérez</i> -----	511
Connecting KOSs and the LOD Cloud <i>Rick Szostak, Andrea Scharnhorst, Wouter Beek, Richard P. Smiraglia</i> -----	521
Converting UDC to BCC: comparative approaches to interdisciplinarity <i>Richard P. Smiraglia, Rick Szostak</i> -----	530

Culturally relevant relationships: publishing and connecting digital objects in collections of archives, libraries, and museums over the Web <i>Carlos H. Marcondes</i> -----	539
Derivative interpretation of biographical sketches (bios) supporting innovative information access <i>Marcia Lei Zeng, Sophy Shu-Jiun Chen</i> -----	549
Extending the scope of library discovery systems via hashtags <i>Louise F. Spiteri</i> -----	557
An FRBR-based approach for transforming MARC records into linked data <i>Ya-Ning Chen</i> -----	565
From uniform identifiers to graphs, from individuals to communities: what we talk about when we talk about linked person data <i>M. Cristina Pattuelli</i> -----	571
Identifying semantic characteristics of user interaction datasets through application of a data analysis <i>Fernando de Assis Rodrigues, Pedro Henrique Santos Bisi, Ricardo César Gonçalves Sant'Ana</i> -----	581
Image organization on the Web: an analysis from the perspective of cultural heritage of rural farms in Brazil <i>Luciana de Souza Gracioso, Leticia Reis da Silveira, Maria da Graça Simões, Luzia Sigoli Fernandes Costa</i> -----	588
Information access in the digital era: document visualization strategy <i>Francisco Carlos Paletta, Armando Malheiro da Silva</i> -----	597
Integrating libraries, archives, museums and art galleries with Linked Data: initiatives study <i>Ana Carolina Simionato, Felipe Augusto Arakaki, Plácida Leopoldina Ventura Amorim da Costa Santos</i> -----	606
Knowledge Organization in a Web collaborative environment <i>Vânia Mara Alves Lima, Cibele de Araújo Camargo Marques dos Santos</i> -----	613
Knowledge Organization Systems and cultural interoperability in open humanitarian settings <i>Quoc-Tan Tran</i> -----	624
Knowledge Organization systems used in European national libraries towards interoperability of the semantic Web <i>Dorota Siwecka</i> -----	633
The Need to interoperate: structural comparison of and methodological guidance on mapping discipline-specific subject authority data to wikidata <i>Andreas Oskar Kempf</i> -----	644

Phenomenon-based vs. disciplinary classification: possibilities for evaluating and for mapping <i>Claudio Gnoli, Andreas Ledl, Ziyong Park, Marcin Trzmielewski</i> -----	653
Relationship status: libraries and linked data in Europe <i>Diane Rasmussen Pennington, Laura Cagnazzo</i> -----	662
The Role of Knowledge Organization tools in open innovation platforms <i>Ricardo Eito-Brun</i> -----	666
Subject access of Mexican television news broadcasts on the Web <i>Silvano Soto-Hernández, Catalina Naumis-Peña</i> -----	674
Tags on healthcare information websites <i>Marit Kristine Ådland, Marianne Lykke</i> -----	684
Towards integrated systems for KOS management, mapping, and access: Coli-conc and its collaborative computer-assisted KOS mapping tool Cocoda <i>Uma Balakrishnan, Jakob Voß, Dagobert Soergel</i> -----	693
Towards the semantic annotation and the prevention of the loss of information of second opinion requests from rural Brazilian primary healthcare providers: the Q-codes use case – a work in progress <i>Melissa P. Resnick, Elena Cardillo, Marc Jamouille, Magdala de Araujo Novaes, Frank S. Shamenek</i> -----	702
Use of conceptual relations for semantic integration of scientific publications and research data <i>Luana Farias Sales, Luís Fernando Sayão</i> -----	711
A Value-based approach to modelling interoperability in Knowledge Organization Systems <i>John Adetunji Adebisi, Babajide Samuel Afolabi, Bernard Ijesunor Akhigbe</i> -----	720
VIAF and OpenCitations: cooperative work as a strategy for information organization in the linked data era <i>Erika Alves dos Santos, Marcos Luiz Mucheroni</i> -----	729

Societal Challenges in Knowledge Organization

Author information for Knowledge Organization in the digital age <i>Wan-Chen Lee</i> -----	739
Challenges of organization and retrieval of photographs on social networks on the Internet <i>Anna Carla Almeida Mariz, Raquel Oliveira Melo, Thales Almeida Mariz</i> -----	746
Challenges to Knowledge Organization in the era of social media: the case of social controversies <i>Nathanaëla Andrianasolo, Adrian-Gabriel Chifu, Sébastien Fournier, Fidelia Ibekwe-SanJuan</i> -----	754

Copyright infringement: between ethical use and legal use of information <i>Arthur Coelho Bezerra, Tatiana Sanches</i> -----	762
Dealing with the paradoxes of customer opinion for effective decision support in churn management <i>Ayodeji O.J Ibitoye, Olufade F.W Onifade, Chika O Yinka-Banjo</i> -----	770
Design Science as a methodology for the development of Knowledge Organization Systems in museological entities <i>Mariana Cantisani Padua, Maria José Vicentini Jorente, Natalia Nakano</i> -----	786
Digital heritage: challenges and opportunities in the access and organisation of digital knowledge in contemporary societies <i>Tiago Trindade Cruz</i> -----	795
Enhancement of digital heritage through digital social networks <i>Camille Rondot, Emmanuelle Chevy-Pébayle</i> -----	804
Epistemology and Ethics of big data <i>Moisés Rockembach, Armando Malheiro da Silva</i> -----	812
Information Design as knowledge technology in the organization of digital information environments on the Zika Virus and its effects <i>Maria José Vicentini Jorente</i> -----	820
Intersectionality and the social construction of Knowledge Organization <i>Maria Aparecida Moura</i> -----	830
Knowledge Organization in the digital age: the complexity of the global labor market <i>Francisco Carlos Paletta, Armando Malheiro da Silva</i> -----	839
Memories in dispute, and reconfigurations of cultural heritage: for an Ethnography of museums <i>Samuel Ayobami Akinruli, Luana Carla Martins Campos Akinruli</i> -----	848
The Mobile phone between challenge and expectations: a potential for information sharing between Algerian breeders and veterinarians <i>Radia Bernaoui, H. Peter Ohly, Kebbab Salim, Mohamed Hassoun</i> -----	856
Negotiating participatory KO in crowdsourcing infrastructures <i>Ina-Maria Jansson</i> -----	863
New ways to produce shared knowledge to improve cooperation in overcoming societal challenges in healthcare: the lever of innovative interface organizations in France <i>Christian Bourret</i> -----	871

The Role of Neuroscience in information and knowledge appropriation <i>António José de Bastos Leite, Francisco Carlos Paletta, Maria Fernanda da Silva Martins, Teresa Silveira</i> -----	880
--	-----

Testing library catalog analysis as a bibliometric indicator for research evaluation in Social Sciences and Humanities <i>Maria Teresa Biagetti, Antonella Iacono, Antonella Trombone</i> -----	892
--	-----

What do museum website users expect from linked open data? <i>Alexandre Fortier, Elaine Ménard</i> -----	900
---	-----

What is happening about KOS in Spain: scientific production analysis, 2000-2017 <i>José A. Moreira-González, Virginia Ortiz-Repiso</i> -----	908
---	-----

Posters

Challenges in management and Knowledge Organization of documental heritage: key factors in the methodology of diligent search for orphan works <i>Rosario Arquero-Avilés, Brenda Siso-Calvo, Gonzalo Marco-Cuenca, Silvia Cobo-Serrano</i> -----	921
---	-----

A Connotative trust-based paradigm to minify societal challenge(s) in Knowledge Organization Systems <i>Ojo Stephen Aderibigbe, Bernard Ijesunor Akhigbe, Babajide Samuel Afolabi, Emmanuel Rotimi Adagunodo</i> -----	924
---	-----

The Contribution of research on information quality to Knowledge Organization <i>Priscila Basto Fagundes, Douglas Dyllon Jeronimo de Macedo, Enrique Muriel Torrado, Adilson Luiz Pinto</i> -----	927
--	-----

Creation of a domain ontology in CIDOC CRM OWL format using heterogeneous textual data related to industrial heritage <i>Eric Kergosien, Kaouther Ben Smida, Rémi Cardon, Natalia Grabar, Mathilde Wybo</i> -----	931
--	-----

Diplomatic forensics: a necessary historical review for the analysis of the born-digital record <i>Juan Bernardo Montoya-Mogollón, Sonia Maria Troitiño</i> -----	937
--	-----

The French military documentary system to anticipate health risk: content and information classification <i>M.Tanti</i> -----	940
--	-----

Gamification as a system for developing knowledge in the classroom: a proposal based on an educational innovation project <i>Rosario Arquero-Avilés, Gonzalo Marco-Cuenca, Alicia Arias-Coello, Brenda Siso-Calvo, Silvia Cobo-Serrano, Carmen Soler-Vaquero</i> -----	945
---	-----

Interdisciplinary research organization: superimposing linked library data, linked research information and research data with interdisciplinary Knowledge Organization <i>Ingo Frank</i> -----	948
Knowledge Organization and information retrieval in institutional repositories <i>Agnes Hajdu Barat</i> -----	951
Knowledge Organization System interoperability: the cogitation of user interfaces for better interactivity <i>Nkechinyere Joy Olawuyi, Bernard Ijesunor Akhigbe, Babajide Samuel Afolabi</i> -----	955
Mapping perspectival ambiguity in Bioethics: revisiting the viewpoint warrant <i>Denis Kos, Sonja Špiranec, Ante Čović</i> -----	959
A Model for decision making (DM) in territory management: social implications <i>María J. López-Huertas, Diego Martín Oliva</i> -----	962
State of the art of organization and administration of libraries in Brazil: preliminary results <i>Jaqueline Santos Barradas, Eliane Cristina Maceió Ferreira</i> -----	966
Web archiving of elections and Brazilian possibilities <i>Moisés Rockembach, Lisiane Braga Ferreira</i> -----	969
 Workshops	
The Politics of classification <i>Robert D. Montoya, Melanie Feinberg, Jonathan Furner, Gregory H. Leazer, Joseph T. Tennis</i> -----	975
Thematic representation and metadata of photographic documents <i>Ana Cristina de Albuquerque, Ana Carolina Simionato</i> -----	979
 List of Contributors and Authors' Index -----	983

Fernando de Assis Rodrigues, Pedro Henrique Santos Bisi, Ricardo César Gonçalves Sant'Ana

Identifying semantic characteristics of user interaction datasets through application of a data analysis

Abstract

In evaluating a decision, any fact analyzed needs to receive inputs from multiple data sources - structuring, integrating, storing, and processing collected data into an output that supports a better understanding of the fact from data, allowing new dimensions of analysis. The goal of this study is to identify the semantic characteristics of data attributes at the moment of collection, from dataset structures found in the data export interfaces of user interaction analysis tools, in Internet communication channels, and in web analytics data tools involved in scientific journal management, through the application of a process of data analysis and data modeling techniques. The research was delimited to exportable datasets available in interfaces from Open Journal Systems, Google Analytics and Search Console, Twitter Analytics, and Facebook Insights. An exploratory analysis approach was adopted to identify characteristics regarding how data are made available and structured in these data resources. Entity-Relationship Modeling concepts were applied to design and store data collected from services, resources, datasets, and attributes. In addition, the collected data was processed into another data structure, adopting the online analytical processing cube as a three-dimensional representation of elements, to facilitate analysis from different perspectives. This data analysis identified semantic dissonances in definitions of entity attributes, which may interfere with the process of developing relationships between attributes from different datasets, reducing the potential of interoperability.

Introduction

The use of data is part of the decision-making process in several fields, such as in education (Ikemoto & Marsh, 2007), industry (Reddy, Srinivasu, Rao and Rikkula 2010), management (Goodwin and Wright 2014), and science (Turban, Aronson and Liang 2004), among others.

In evaluating a decision, any fact analyzed needs to receive inputs from multiple data sources – structuring, integrating, storing, and processing the collected data into an output that supports a better understanding of the fact from data, allowing new dimensions or perspectives of analysis (Inmon 2005; Kimball and Ross 2011; Reddy *et al.* 2010; Turban *et al.* 2004).

For example, an evaluation of interactions between users and scientific contents in a publisher's web domain may be analyzed by service holders from the outputs generated in a process of collecting data regarding users' interactions with their communication channels, structured into a data warehouse: a “[...] subject-oriented, integrated, time-variant, non-normalized, non-volatile collections of data that support analytical decision-making” with “[...] access to all information relevant to the organization, which may come from many different sources, both internal and external” (Turban *et al.* 2004, p. 236).

However, if data are analyzed as a set of elements formed by the triad of entity, attribute, and value (Santos and Sant'Ana 2015), this means using aggregated information in these elements to assure minimal semantics to understand what is available, particularly in regard to steps in obtaining data collected from data sources (Sant'Ana 2016; Turban *et al.* 2004).

This effort to bind information in these data elements seeks to minimize semantic dissonance between data, at the moment of data collection (Berg 2015; Rathod 2006; Ross Parry, Nick Poole and Jon Pratty 2008) – the research problem of this study.

Our aim is to identify the semantic characteristics of data attributes at the moment of collection, from dataset structures found in the data export interfaces of user interaction analysis tools, in Internet communication channels, and in web analytics data tools involved in scientific journal management, through the application of a process of data analysis and data modeling techniques.

The research was delimited to exportable dataset structures, found in journal publishing systems, online social network statistics, search engines, and web analytics tools.

The sample was restricted to dataset structures available in reports from Open Journal Systems¹, Google Analytics², Google Search Console³, Twitter Analytics⁴, and Facebook Insights⁵. These resources did not present any version control numbering on their interfaces, with the exception of Open Journal Systems (version 2.6). The data was collected in September 2017 from "Electronic Journal Digital Skills for Family Farming (RECoDAF)" accounts.

Methodology

An exploratory analysis approach was adopted to identify characteristics regarding how data are made available and structured in these data resources, contemplating a systematic description process for information from datasets, entities, and attributes related to interaction between users and communications channels of a scientific journal.

A total of 255 exportable datasets were found, distributed in 5 file formats: Comma-Separated Values (CSV) (82 datasets), Google Docs Spreadsheet File Format (69 datasets), Microsoft Office Open XML Format Spreadsheet file (XLSX) (50 datasets), Portable Document Format (PDF) (50 datasets), and Microsoft Excel Binary File Format (XLS) (3 datasets).

¹ Open Journal Systems is an open-source software developed by Public Knowledge Project, under GNU General Public License.

² Google Analytics is a web analytics service by Google LLC.

³ Google Search Console (formerly known as Google Webmaster Tools) is a web service by Google LLC.

⁴ Twitter Analytics is a web analytics service by Twitter, Inc.

⁵ Facebook Insights is a web analytics service by Facebook, Inc.

The 82 CSV datasets are distributed on 5 services, 50 retrieved from Google Analytics, 20 from Google Search, 7 from Open Journal Systems, 3 from Facebook Insights, and 2 from Twitter Analytics.

Except for CSV, all other file formats were discarded. The CSV is the only format available in analyzed data sources that is an Internet mime-type, an open file format (Shafranovich, 2005), an Internet tabular data model (Tennison, Kellogg and Herman 2015), and machine-readable by all well-known programming languages (Lebo and Williams 2010). Moreover, CSV is the only format that appears as an export option in all the interfaces analyzed.

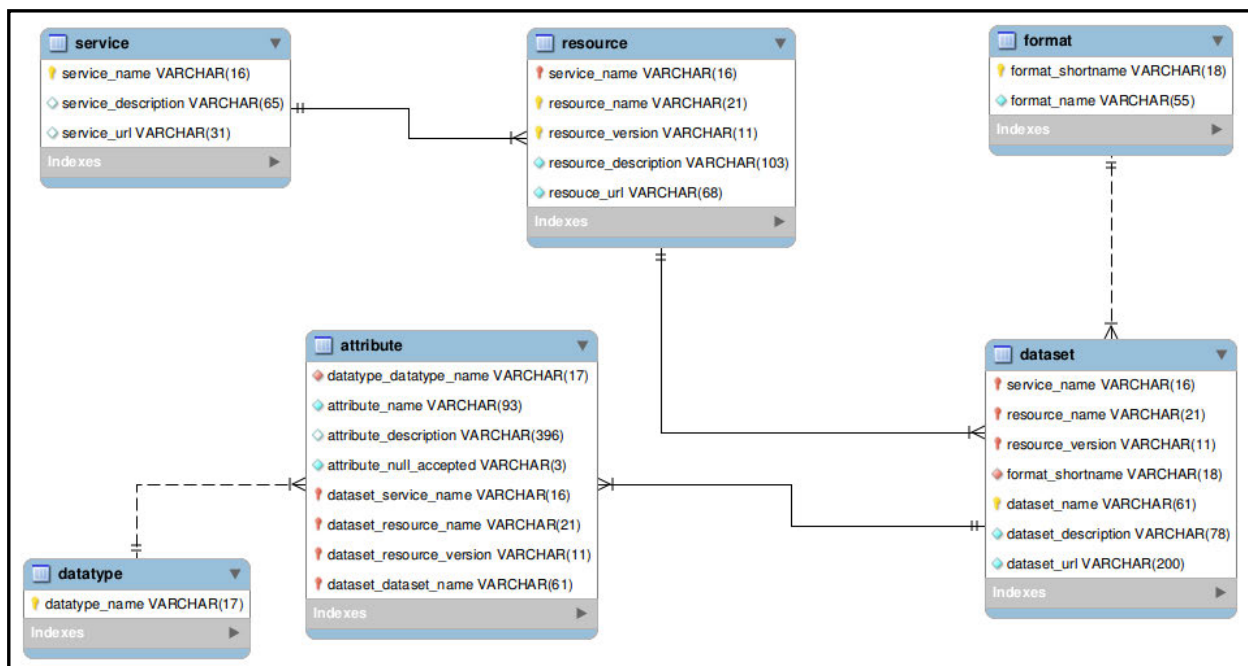
Data analysis

In order to systematize the data analysis, concepts from Entity-Relationship (ER) Modeling (Silberschatz, Korth and Sudarshan 2011) were applied; using the set of conventions from ER "[...] to assist in databases design processes" (Date 2016, p. 64).

An ER model was developed (Figure 1), designed to store data collected from (i) services, (ii) resources available in each service, (iii) datasets available in each resource, and (iv) attributes available in each dataset.

In addition, two tables were developed to store information about controlled vocabularies applied in datasets and attributes, in order to control the set of available formats and data types in these elements (Date 2016, p. 228).

Figure 1: Diagram of Entity-Relationship model developed for data collecting



As a first step, the ER structure was applied in an Open Document Spreadsheet file (ODS) to store data collected in this study⁶, which was able to identify in the 82 datasets a total of 2,280 attributes, with a subset of 1,342 distinct attribute labels.

The second step was to convert the ODS file to a CSV file and upload and import it into a database management system (DBMS) for subsequent data analysis, with tables and columns representing ER model entities and attributes, respectively.

A script for Python programming language was developed to assist the processing and reordering of data uploaded to tables in the DBMS into a second data structure, adopting the online analytical processing cube⁷ as a three-dimensional representation of services (s), datasets entities (e), and attributes (a), acting as perspectives of analysis (Gray, Bosworth, Lyaman and Pirahesh 1996; Inmon 2005; Kimball and Ross 2011).

The collected data was reordered to OLAP cube dimensions by concepts derived from the pivot table process (Cornell 2005).

To evaluate an OLAP fact, we intended to observe the intersections of the OLAP cube to determine the characteristics shared internally and externally by services, entities and attributes that may affect semantic issues of data collection.

Results

The data analysis identified several attributes with label names composed by filtering, grouping or sorting specifications as a part of text, a pattern followed only by online social network statistical data export tools. This leads to an increase of complexity involved in how to interpret those attributes and label inherent characteristics as values, using fully or semi-automated data collecting algorithms.

In this scenario, it is possible to determine that an entity (e_x) may have two attributes (a_x and a_y) sharing the same semantics (S), even when both attributes show distinct text labels in data collecting, expressed by the formula:

$$S(e_x, a_x) = S(e_x, a_y)$$

An example that fits in this formula are two attributes (a_x and a_y) from a unique entity (e_x), with filtering specifications as a part of text labels, representing the total of people who engaged with the journal content on an online social network profile (S) by geographic area.

From another perspective, results from data analysis identified a larger set of attributes that do not relate to any description of its content, formed by 88.69% of available attributes, which means that label is the only explicit information on those attributes available at the moment of data collection.

⁶ The data collected in this research are available at http://dadosabertos.info/data/collection_recodaf_2017.

⁷ Also known as OLAP cube.

Furthermore, all attributes that share equal label names are part of a subset of attributes that do not have any description. This is critical in a data collecting process, primarily because this subset of attributes plays a significant role in the interoperability of data, inherently capable of being part of the set of potential primary keys with unique value restrictions, helping to build relationships between data sources, or determining geographic, temporal or linguistic aspects of the content itself.

This absence of semantics, with the exception of the availability of text labels, does not ensure that attributes of two distinct entities (e_x and e_y) that share equal labels (a_x) will, consequently, share the same formal semantics (S) in data collection by external agents, expressed by the formula:

$$S(e_x, a_x) \neq S(e_y, a_x)$$

That effect requires external teams to interpret the semantics of these elements locally, aided by their skills or previous knowledge.

For example, two attributes that share equal text labels (a_x) from distinct entities (b_x and b_y), without proper description of their content, may require interpretation of formats, data types, primary keys, unique restrictions, and controlled vocabularies applied, increasing the risk of wrong interpretations of values, thus preventing data collection teams from understanding that attributes may share the same text labels but not the same semantics (S).

Conclusions

Data analysis helped to identify the critical points related to the adherence of descriptive elements in the datasets analyzed, especially the lack of descriptive elements in the data collection process when triggered through the available export interfaces.

To reduce this dissonance between attributes, export interfaces could provide more semantic information bound to datasets. This information may be fundamental to interpret data available from different sources. Therefore, one action to reduce semantic dissonances between attributes is the enhancement of text labeling rules, including the use of controlled vocabularies and restriction clauses.

Moreover, the semantic dissonances in these entities may interfere with the development process of relationships between attributes from different datasets, thereby reducing the potential for interoperability.

References

- Berg, O. (2015). *Collaborating in a social era: ideas, insights and models that inspire new ways of thinking about collaboration*. Göteborg: Intranätverk.
- Cornell, P. (2005). *A Complete guide to PivotTables: a visual approach*. Berkeley, CA: New York: Apress; Distributed to the Book trade in the United States by Springer-Verlag.
- Date, C. J. (2016). *The New relational database dictionary: a comprehensive glossary of concepts arising in connection with the relational model of data, with definitions and illustrative examples: [terms, concepts, and examples]*. Sebastopol, CA: O'Reilly.
- Goodwin, P., & Wright, G. (2014). *Decision analysis for management judgment* (5th ed.). Hoboken, New Jersey: Wiley.
- Gray, J., Bosworth, A., Lyaman, A., & Pirahesh, H. (1996). Data cube: a relational aggregation operator generalizing GROUP-BY, CROSS-TAB, and SUB-TOTALS (p. 152-159). IEEE Comput. Soc. Press. Retrieved from: <https://doi.org/10.1109/ICDE.1996.492099>.
- Ikemoto, G. S., & Marsh, J. A. (2007). Cutting Through the “Data-Driven” Mantra: different conceptions of data-driven decision making. *Yearbook of the National Society for the Study of Education*, 106(1): 105-131. Retrieved from: <https://doi.org/10.1111/j.1744-7984.2007.00099.x>.
- Inmon, W. H. (2005). *Building the data warehouse* (4th ed). Indianapolis, Ind: Wiley.
- Kimball, R., & Ross, M. (2011). *The Data Warehouse Toolkit: the Complete Guide to Dimensional Modeling*. New York, United States of America: John Wiley & Sons. Retrieved from: <http://nbn-resolving.de/urn:nbn:de:101:1-2014122311140>.
- Lebo, T., & Williams, G. T. (2010). Converting governmental datasets into linked data. In *Proceedings of the 6th International Conference on Semantic Systems*. Graz, Austria: ACM Press. Retrieved from: <https://doi.org/10.1145/1839707.1839755>.
- Rathod, A. (2006). *A Messaging system to handle semantic dissonance*. New York: Rochester Institute of Technology. (Thesis). Retrieved from: <http://scholarworks.rit.edu/cgi/viewcontent.cgi?article=1668&context=theses>.
- Reddy, G. S., Srinivasu, R., Rao, M. P. C., & Rikkula, S. R. (2010). Data warehousing, data mining, OLAP, OLTP technologies are essential elements to support decision-making process in industries. *International Journal on Computer Science and Engineering*, 2(9): 2.865-2.873.
- Ross Parry, Nick Poole, & Jon Pratty (2008). Semantic Dissonance: do we need (and do we understand) the semantic Web? In *Toronto: Archives & Museum Informatics*. Retrieved from: <http://www.archimuse.com/mw2008/papers/parry/parry.html>.
- Sant’Ana, R. C. G. (2016). Ciclo de vida dos dados: uma perspectiva a partir da Ciência da Informação. *Informação & Informação*, 21(2): 116. Retrieved from: <https://doi.org/10.5433/1981-8920.2016v21n2p116>.
- Santos, P. L. V. A. da C., & Sant’Ana, R. C. G. (2015). Dado e granularidade na perspectiva da Informação e Tecnologia: uma interpretação pela Ciência da Informação. *Ciência da Informação*, 42(2): 11.

- Shafranovich, Y. (2005). Common Format and MIME Type for Comma-Separated Values (CSV) Files. *The Internet Society*. Retrieved from: <https://tools.ietf.org/html/rfc4180>.
- Silberschatz, A., Korth, H. F., & Sudarshan, S. (2011). *Database system concepts* (6th ed.). New York: McGraw-Hill.
- Tennison, J., Kellogg, G., & Herman, I. (2015, December 17). *Model for Tabular Data and Metadata on the Web*. (J. Tennison & G. Kellogg, ed.). World Wide Web Consortium. Retrieved from: <https://www.w3.org/TR/tabular-data-model/>.
- Turban, E., Aronson, J. E., & Liang, T.-P. (2004). *Decision Support Systems and Intelligent Systems* (7th ed.). Upper Saddle River, NJ, USA: Prentice-Hall, Inc.