ADVANCES IN KNOWLEDGE ORGANIZATION VOL. 16

INTERNATIONAL SOCIETY FOR KNOWLEDGE ORGANIZATION [ISKO]

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

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.

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

Interoperability towards Information Access

Societal Challenges in Knowledge Organization

Posters

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 Systems[1], Google Analytics[2], Google Search Console[3], Twitter Analytics[4], and Facebook Insights[5]. 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.

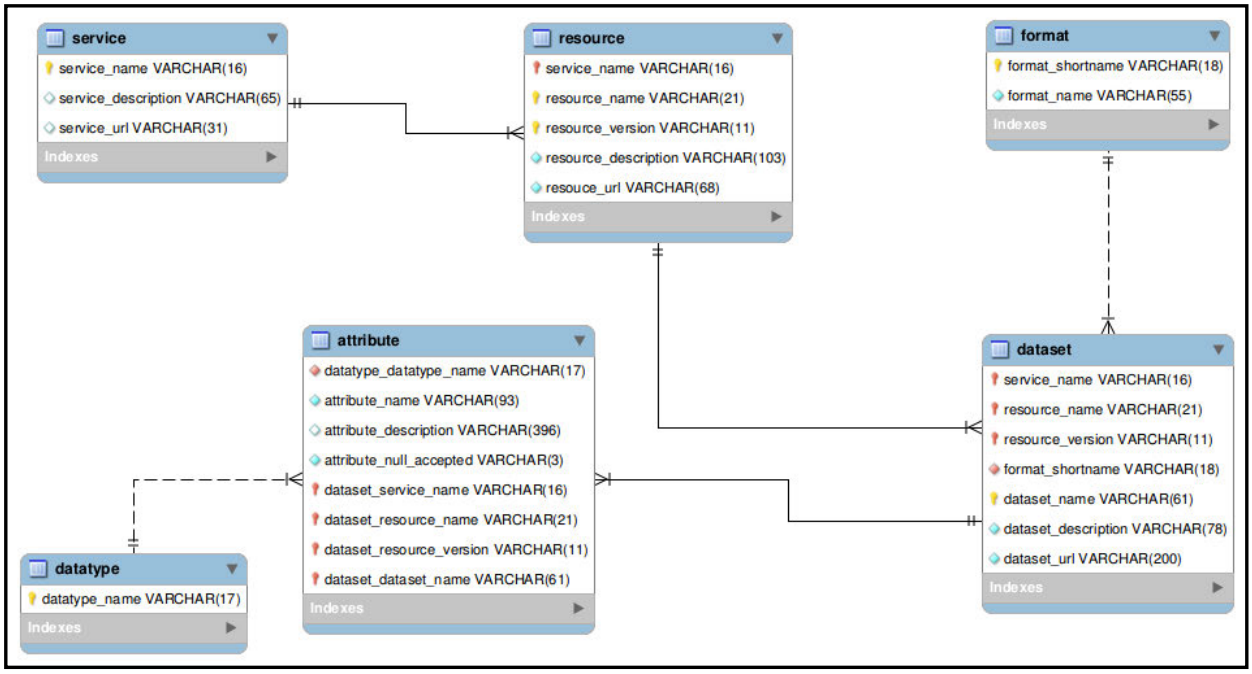
Data analysis

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 study[6], 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 cube[7] 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.