Contents

Vanda Broughton.
Obituary:
Emeritus Professor Ia McIlwaine:
An Appreciation ................................................................. 573

Special Issue:
Best papers from NASKO, ISKO-UK,
ISK0-France, ISKO-Brazil 2019

Articles

Heather Moulaison Sandy and Andrew Dillon.
Mapping the KO Community................................................ 578

Elliott Hauser and Joseph T. Tennis.
Episemantics: Aboutness as Aroundness ....................... 590

Vanda Broughton.
The Respective Roles of Intellectual Creativity
and Automation in Representing Diversity:
Human and Machine Generated Bias................................. 596

Shu-Jiun Chen.
Semantic Enrichment of Linked Personal Authority
Data: A Case Study of Elites in Late Imperial China ........... 607

Viviane Clavier.
Knowledge Organization, Data and Algorithms:
The New Era of Visual Representations............................. 615

Joachim Schöpfel, Dominic Varwe, Hélène Prost and
Antonella Zane.
Data Papers as a New Form of Knowledge
Organization in the Field of Research Data ...................... 622

Pablo Gomes and Maria Guiomar da Cunha Prata.
Knowledge Organization from a Social Perspective:
Thesauri and the Commitment to Cultural Diversity .......... 639

Mario Bartié.
Towards a General Conception of Warrants:
First Notes........................................................................... 647

Research Trajectories in Knowledge Organization

Barbara H. Kwaśniki.
Changing Perspectives on Classification as a
Knowledge-Representation Process.................................. 656

Sixth Annual “Best Article in KO Award”
for Volume 45 (2018) .......................................................... 668

Books Recently Published.................................................... 669
Obituary.

Emeritus Professor Ia McIlwaine:
An Appreciation

Professor Ia McIlwaine was born Ia Cecilia Thorold on 20 April 1935, the only daughter of Michael Thorold, a clergyman in the Church of England, and Dorothy Henfrey; she had two younger brothers. Her unusual Christian name, which often caused confusion for correspondents, is Cornish in origin, the nominative form of what we may be more familiar with in the place name St. Ives.

She was a pupil at Bath High School, and, after leaving school, went up to Bedford College, London to read Classics, a discipline in which she maintained a lifelong interest. In 1957-58 she studied for the Graduate Diploma in Librarianship at University College London, the beginning of a long and distinguished association with that institution. After a five year period as Assistant Librarian with Westminster City Libraries, during which time she was awarded Fellowship of the Library Association, she was appointed to the post of Lecturer in the School of Library & Archive Studies at UCL. A major part of her role there was to teach classification and indexing, the subject field which would become her primary research area.

She progressed steadily through the academic ranks, being promoted to Senior Lecturer in 1985 and Reader in Classification and Indexing 1995, at which time she also took on the Directorship of the School, now the School of Library, Archive & Information Studies. In 1997 she was honoured with a Chair of Library & Information Studies. She continued the strong tradition of classification and indexing work in the School, which had begun with Charles Berwick Sayers, whose pupil Ranganathan was in the 1920s. At a time of waning interest in classification in most UK library schools, she kept it solidly and centrally on the curriculum at UCL, and sustained it as a distinctive feature of the UCL department with a programme of international events and a lively group of research students; some of her doctoral students now occupy leading roles in the world of library and information science in general, and classification in particular. As a student myself in the early 1970s I found the classification element of the course the most appealing and intellectually engaging, and it led me into the most rewarding career which I would not have enjoyed without her original enthusiasm and expertise.

Although her academic work was primarily focused on classification, where most of her publications are to be found, she had a broader interest in subject work and bibliography generally, and in bibliographic control. She co-authored the book *Introduction to Subject Study*, and edited a number of conference and collected papers including *Standards for the International Exchange of Bibliographic Information, Subject Retrieval in a Networked Environment*, and *Knowledge Organization and the Global Information Society*. Her reputation as an editor led to the role of series editor for Saur’s (later de Gruyter) substantial *Introduction to Information Sources*, which documented the bibliography of a wide number of disciplines and formats. On the classification front, she was an active member of the International Society for Knowledge Organization, being President from 2001-2005, as well as a longstanding member of the Scientific Advisory Committee, and the Editorial Board of its journal *Knowledge Organization*. She was also, in its latter years, Secretary of the UK Classification Research Group. In addition, she was a

DOI: 10.5771/0943-7444-2019-8-573
member of the British Standards Institution Committee on Indexing, an observer on the Committee of the Bliss Classification Association, and served on the British Committee for the Dewey Decimal Classification.

In what was probably her major contribution to the field, she took over in 1993 as Editor-in-Chief of the Universal Decimal Classification, leading a major programme of revision of this large international system. A major part of the revision work under her Editorship was the introduction of a more rigorous analytical and facetted approach to the classification as a whole, which resulted in a number of radically revised classes, as well as a systematic pruning and rationalisation of the scheme overall. To this end, she engaged the help of a number of colleagues and contributors, with whose assistance several substantial revisions of main classes were achieved. She undertook much of this painstaking and time-consuming work herself, producing new schedules for photography and large parts of the auxiliary table for “place,” as well as a major overhaul of the medicine class, carried out in association with Nancy Williamson.

She gave much thought to the problems of maintenance and revision of large classification systems, particularly against the background of transition to an online environment. Within that context, another innovative feature of her time as editor of the UDC was her fostering of a more cooperative and collaborative view of classification revision. She worked closely with the editors of both the Dewey Decimal Classification, and the Bliss Bibliographic Classification, both to share the effort of maintaining these large systems, and also to promote consistency in the way they represented subject content.

Alongside her work in classification and knowledge organisation she retained her interest in the classics. Her PhD work, published as Herculaneum: A Guide to Printed Sources, was also carried out at UCL while she was a member of staff, and was awarded the prestigious Dunn & Wilson prize. This interest was picked up again in retirement, when she produced a supplement to her thesis for a new publication from Bibliopolis for Centro Internazionale per lo studio dei papiri ercolanesi. For this scholarly work she was elected to a Fellowship of the Society of Antiquaries, an honour which particularly pleased her.

In addition to her academic work, Ia was a powerful advocate for the profession at both national and international level. In 1998 she was recipient of the United Kingdom Library Association Centenary Medal, one of a hundred members of the Association so honoured, for her “services to the profession,” largely an acknowledgement of a long career spent in education for librarianship. She had an equally longstanding commitment to the International Federation of Library Associations, and served on its Governing Board from 1993-2003, chairing the IFLA Professional Committee from 2001-2003. She was also a member and office holder in the Section for Classification and Indexing and the Division of Knowledge Management, and in 2005 was awarded the IFLA Medal, “for distinguished services to IFLA.”

Much of this activity continued into her early years of retirement, but her time was increasingly spent at the family cottage in Norfolk, and later at a larger house where she could indulge her love of gardening. After a period of ill health she died on 24 August 2019 from complications following pneumonia. In 1966 she had married her fellow lecturer, John McIlwaine, who was her colleague and her companion for 53 years. He survives her, together with their daughter Anne who followed her parents into the world of libraries.

Ia was tireless in her work for the School of Librarianship, for University College, and for the wider world of librarianship, never shirking a difficult situation, and always going the extra mile. It was often her personal involvement and attention to detail that ensured the success of any number of activities, both at home and abroad, including the hosting of several international conferences on classification and information retrieval. A sometimes daunting manner hid a well of personal kindness, and a considerable sense of responsibility. She was a formidable administrator, but also a great enabler of younger colleagues and associates who widely acknowledge the role she played in encouraging and advancing their roles in the profession.

It was a not uncommon occurrence for some overseas visitor to stroll into the School Office at UCL, uninvited and unannounced (and usually on a Friday afternoon), confident that she would welcome them, which she invariably did. A vast network of connections across the libraries of the world bears testimony to the friendship and respect she was accorded by her fellow practitioners, as well as her academic contacts in all countries. Her part in the international scene can hardly be exaggerated, and she will be much missed by her friends and colleagues across the globe.

Vanda Broughton
Emeritus Professor of Library & Information Studies, University College London
<v.broughton@ucl.ac.uk>

Select bibliography

Librarianship and bibliography generally


**Classification and information retrieval**


**Universal Decimal Classification**


the Introduction of Greater Facet Analysis into the UDC." Extensions & Corrections to the UDC 15: 31–44.


Herculaneum


University College London School of Librarianship


Mapping the KO Community†

Heather Moulaison Sandy* and Andrew Dillon**

*University of Missouri, iSchool, 303 Townsend Hall, Columbia, MO 65211, U.S.A., <moulaisonhe@missouri.edu>
**University of Texas, School of Information, 1616 Guadalupe St, Suite 5-202, Austin TX, 78701, <adillon@ischool.utexas.edu>

Heather Moulaison Sandy is Associate Professor at the iSchool at the University of Missouri. Her research focuses primarily on the intersection of the organization of information and technology and includes the study of issues pertaining to metadata, standards, and scholarly communication. An ardent Francophile, Moulaison Sandy is also interested in international aspects of access to information.

Andrew Dillon is the V. M. Daniel Professor of Information at the School of Information, UT-Austin where he also holds appointments in psychology and info, risk and operations management. He studies the psychological dynamics of information use and design. Dillon is co-editor of the journal Information & Culture and serves on the editorial boards of the Journal of Documentation and Interacting with Computers.


Abstract: Knowledge organization (KO) is considered a distinctive disciplinary focus of information science, with strong connections to other intellectual domains such as philosophy, computer science, psychology, sociology, and more. Given its inherent interdisciplinarity, we ask what might a map of the physical, cultural, and intellectual geography of the KO community look like? Who is participating in this discipline's scholarly discussion, and from what locations, both geographically and intellectually? Using the unit of authorship in the journal Knowledge Organization, where is the nexus of KO activity and what patterns of authorship can be identified? Cultural characteristics were applied as a lens to explore who is and is not participating in the international conversation about KO. World Bank GNI per capita estimates were used to compare relative wealth of countries and Hofstede's Individualism dimension was identified as a way of understanding attributes of countries whose scholars are participating in this dialog. Descriptive statistics were generated through Excel, and data visualizations were rendered through Tableau Public and TagCrowd. The current project offers one method for examining an international and interdisciplinary field of study but also suggests potential for analyzing other interdisciplinary areas within the larger discipline of information science.

Received: 5 September 2019; Revised: 9 September 2019; Accepted: 26 October 2019

Keywords: authors, knowledge organization, articles, research


1.0 Introduction

Knowledge organization (KO) is sometimes narrowly conceived as a concern of library and information science professionals, but even a quick examination at the affiliations of authors publishing in the field reveals that other intellectual domains such as philosophy, computer science, business, psychology, linguistics, sociology, and more contribute to and find value in its study. The subject matter of KO embraces fundamental questions of what constitutes knowledge as well as practical concerns of how to represent and enable access for others. Accordingly, it can be difficult to characterize and understand the domain of KO or to position it intellectually both academically and professionally.

Academic journals provide a forum for the exchange of new knowledge in a discipline and serve as a record of the contributions made to a domain or field across time. As such, a scholarly journal serves to validate research, and by extension, helps to shape the legitimacy of a field of inquiry. Long-standing journals in a domain are considered to provide a measure of prestige for authors as well as an identity for a discipline. New areas of enquiry or research involving non-traditional methods often face a challenge gaining a foothold in academia until a suitable peer-reviewed outlet such as an academic journal or high prestige
By virtue of this type of gatekeeping role, academic journals can provide useful indices of the development of a domain and the research participants within it. Consequently, it is possible to use the back issues of a journal as a test base for examining the emergence, duration, and impact of ideas within a field, as well as the productivity of key scholars. Darmani, Dwaikat and Portilla (2013), for example, analyzed ten years of contributions to the Journal of Creative Innovation and Management to shed light on how the field of innovation management is evolving over time and to determine the geographical make up of scholarship in the domain. By characterizing author geography, publication trends, and recurring themes across a decade, they provided evidence of the diminishing occurrence of single-author papers, the recent growth of scholarship from emerging economies, and the dominance of leadership as a primary research emphasis. Similarly, Wäid, du Preez and Wallström (2012) performed an analysis of Marketing Intelligence and Planning to identify major author patterns and content trends in the field of marketing, highlighting the location of key authors and the productivity of regions in generating new scholarship. Such work can prove useful in encouraging a shared perspective and identifying areas of need within a subject or discipline. The present paper represents an attempt at a similar analysis in the domain of KO.

1.1 Efforts to assess the KO community

Previous studies of KO have addressed questions relating to the field’s geographic reach and intellectual focus. Zhao and Wei (2017), for example, study collaborations among Chinese authors in KO from 1992 through 2016. In examining 1,298 articles with Chinese authors published in Web of Science Core Collection KO journals, they find an increase in collaborations over the period of study, including in international collaborations (from 50% in 1992 to 92.53% in 2016). Likewise, Smiraglia (2015) investigates the field to evaluate the work being done in the area of domain analysis, a unique area of study covered in KO. Beyond KO, scholars in LIS have studied the international contributions to the Journal of the American Society for Information Science and Technology (JASIST) and in the Journal of Documentation (He and Spink 2002) over a fifty-year period at the time when electronic journals were changing the scholarly communication landscape. Analyzing first author affiliations only, these authors report that international contributions increased over the time of study (1950-1999) for both journals. The extent to which KO mirrors the broader discipline or represents a distinct area with unique or distinctive scholarly characteristics in its corpus remains an open question.

1.2 Metrics to assess countries, comparatively

Broad estimates of global expenditure on research suggest where scholarly efforts are most actively pursued, and it is perhaps not surprising that in 2017 the US and Europe accounted for over 45% of annual spending on research and development, with China accounting for a further 22% (Statista 2019). These proportions correlate with the existence and growth of universities globally, though the US continues to dominate regional presence within top research university rankings. Domain or disciplinary differences, though more difficult to determine, also exist and are likely to reflect national and political emphases on research. Chinese universities, for example, are becoming highly ranked in engineering and computer science but less so on liberal arts, which remain dominated by US and European, particularly British, institutions.

Global rankings and expenditures are somewhat limited measures, and we recognize that scholars can, depending on their circumstances, be mobile, gravitating toward and succeeding at institutions that allow for them to investigate questions of interest using the methods that are most applicable. Further, we must acknowledge that scholarship in different countries varies in its reward and recognition, and political and economic support from the public and private sectors. Given the range and the regional differences in support and emphasis for particular research, it is interesting to consider where KO scholarship is situated and how it is distributed and enacted globally.

A number of metrics are available to assess cultural differences, the best-known being those put forth by Hofstede, Hofstede and Minkov (2005). Their metrics, derived from large-scale and long-term surveys, outline six dimensions of culture and profile countries and regions based on their scores across these dimensions. As imperfect as these metrics may be, they have become widely used in business and research and offer a starting point for comparing cultures internationally. In particular, the individualism vs. collectivism dimension has the potential to provide insight into the collaborative nature of scholarship and the writing process around the world. We might expect, for example, that cultures differing on this dimension also manifest distinctive publication styles in terms of single-authored or collaborative articles. Further, we might anticipate that KO, with its interrogation of knowledge structures and authority, might be impacted by cultural distinctions based on power distance or uncertainty avoidance.

Another metric, put forth by the World Bank, assesses relative wealth of a country’s citizens by calculating the gross national income (GNI) of the country on a per-capita basis. Limited by virtue of reducing entire populations to a single measure of income, these numbers might provide a basis for comparison and, in conjunction with Hofstede et al.’s
dimensions. Offer one other gross index to help us better understand what we might term the cultural climate of scholarship.

1.3 Mapping the KO community authors

For more than forty years, the journal Knowledge Organization has served as a primary venue for research and discourse in the field. As such, the journal contains the richest record of the discipline’s content, contributors, and trends and is explored here to provide us with a database of research activities in the field. Using the unit of “authorship,” we seek to identify what countries appear as a nexus of KO activity and what patterns of authorship (and co-authorship) can be found in these data? We wish to characterize the KO community of researchers as it has emerged on empirical grounds to better understand how this area is evolving and how it is positioned intellectually.

To begin to explore these questions along with the cultural and disciplinary factors influencing the domain, this research paper maps the geography of Knowledge Organization authorship. The current project explores a method for analyzing an international and interdisciplinary field of study that we hope might prove useful not just for KO but for other areas of the information discipline in both standalone and comparative studies.

2.0 Method

To assess the question of authorship by nationality based on institutional affiliation, all scholarly articles published in Knowledge Organization from 2009 to 2018 inclusive were examined. New articles that presented research including research articles and revised conference proceedings were considered scholarly and were retained for analysis. For this project, scholarly articles retained included articles labeled “peer reviewed” and research articles that expand on peer-reviewed conference proceedings (usually indicated in the TOC as “Selected Papers from the X Conference” – N.B. these tend to be grouped geographically by ISKO chapter, which affects the mapping of authorship in a way that should be acknowledged. These are nonetheless part of the scholarly record produced by Knowledge Organization, so excluding them would be a mistake). Finally, “Reviews of Concepts” in Knowledge Organization were retained. Editorials, features, brief communications, discussions such as the “Forum: The Philosophy of Classification,” “Classification Research,” “Research Trajectories,” conference reports, “ISK0 News,” book reviews, introductions to special issues, festschrift articles reviewing the life of honorees, and reprintings of previously published articles were not retained for inclusion. Editorials, book reviews, and reprints of seminal articles were also excluded, but any of these could be further analyzed later. Using the individual author as the primary unit of analysis, each contributor to the publication of a scholarly article in Knowledge Organization was identified, and his or her name, institution, school, department, or unit if applicable, the country of the institution, and the total number of co-authors on the article were retained in Excel.

Hofstede’s individualism-collectivism dimension was applied to the data set as a way of understanding relative attributes of countries whose scholars are participating in this dialog. World Bank GNI per capita estimates (https://data.worldbank.org/indicator/ny.gnp.pcap.ppp.cd) in US dollars for 2017 were used to compare relative wealth of countries. Because of our assumptions about the mobile nature of academics and the observation that English has become the lingua franca in scholarly communication, no attempt to understand authors’ country of origin, languages spoken, or educational background was made. Descriptive statistics were generated through Excel; more complex data visualizations were rendered through Tableau Public and TagCrowd.

3.0 Results and discussion

For this project, 362 scholarly articles, with 632 individual statements of authors, were coded for analysis and description. In the first instance, we examined publication rates over time and determined that over the last ten years, there has been almost a doubling of published papers in Knowledge Organization (see Figure 1), though this might reflect exceptional years 2016-2017. Nevertheless, the general trend is positive with increasing number of papers published in Knowledge Organization over time.

A total of 466 unique authors contributed to the articles, with the majority (n=384) of authors contributing to one article, and a minority (n=82) contributing to two or more articles (see Figure 2). What this means for Knowledge Organization as a scholarly venue is not obvious. This might reflect the increasing breadth of new authors publishing in Knowledge Organization or it could be the case of scholars just publishing once here and moving on or not publishing further (in the case of students who publish with professors but then pursue professional careers elsewhere). This is one question that might be usefully pursued over time.

In terms of individual author productivity, twelve authors published four or more scholarly articles over the ten-year period (see Table 1). While traditional author impact and productivity measures are not the focus of this work, it is interesting to note that these twelve individuals’ contributions represent roughly 24% of the journal’s total output. Without comparative data from other fields it is hard to draw conclusions here but at first glance, this proportion of contributions from a rather small set of schol-
ars might be indicative of an emerging rather than a mature field and is likely of some interest to those involved in promotion and tenure discussions.

Authors were affiliated with institutions located in thirty-nine countries. See Figure 3 for a breakdown of the number of authors from Algeria to Singapore by year. This suggests that KO scholarship is indeed global. As expected, the most productive scholars shown above (Table 1) are generally from the countries with the highest representation over time, including the United States, Canada, Brazil, and Denmark.

With authors as the unit of analysis, entries for each author responsible for the scholarly articles studied were coded separately. Figure 4 maps the contributions of these authors, by entry for author. Darker blue countries had higher numbers of total author contributions during the
A ten-year period of study with the largest number of scholarly article authors coming from the United States (n=137) and Brazil (n=105).

The progression over time of international authorship can be seen in Figure 5 (interactive version available online). Also visible is the publication of the revised proceedings of the various biennial ISKO chapter meetings (featured chapters include ISKO France's 2017 conference (2017), ISKO-UK's 2017 conference (2017), ISKO-Brazil's 2017 conference (2017), ISKO-Italy's 2017 conference (2017), ISKO-Brazil's 2015 conference (2016), ISKO Spain-Portugal's 2015 conference (2016), ISKO-Canada/US's 2015 conference (2015), ISKO-Brazil's 2013 conference (2014), ISKO Spain and Portugal's 2013 conference (2014), German ISKO's 2013 conference (2013), ISKO Italy's 2011 conference (2012), ISKO-France's 2011 conference (2012), and others). The biennial international ISKO conference has also been represented. For example, the ISKO Conference 2016 (2016) was also featured.

Table 1. Individual authors contributing four or more scholarly articles to Knowledge Organization, 2009-2018 and their country and school/department affiliations.

<table>
<thead>
<tr>
<th>Author</th>
<th>Country</th>
<th>School or department affiliation</th>
<th>Articles contributed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Birger Hjørland</td>
<td>Denmark</td>
<td>Department of Information Studies</td>
<td>14</td>
</tr>
<tr>
<td>Daniel Martínez-Ávila</td>
<td>Brazil</td>
<td>Department of Information</td>
<td>12</td>
</tr>
<tr>
<td>Claudio Gnoli</td>
<td>Italy</td>
<td>Library</td>
<td>7</td>
</tr>
<tr>
<td>José Augusto Chaves Guimarães</td>
<td>Brazil</td>
<td>Graduate School of Information Science</td>
<td>7</td>
</tr>
<tr>
<td>Richard P. Smiraglia</td>
<td>USA</td>
<td>School of Information Studies, Knowledge Organization Research Group</td>
<td>7</td>
</tr>
<tr>
<td>Elaine Ménard</td>
<td>Canada</td>
<td>School of Information Studies</td>
<td>6</td>
</tr>
<tr>
<td>Joseph T. Tennis</td>
<td>USA</td>
<td>Information School</td>
<td>6</td>
</tr>
<tr>
<td>Margaret E. I. Kipp</td>
<td>USA</td>
<td>School of Information Studies</td>
<td>6</td>
</tr>
<tr>
<td>Melodie J. Fox</td>
<td>USA</td>
<td>School of Information studies</td>
<td>6</td>
</tr>
<tr>
<td>Rick Szostak</td>
<td>Canada</td>
<td>Department of Economics</td>
<td>6</td>
</tr>
<tr>
<td>Fabio Assis Pinho</td>
<td>Brazil</td>
<td>Department of Information Science</td>
<td>5</td>
</tr>
<tr>
<td>Patrick Keilty</td>
<td>Canada</td>
<td>Faculty of Information</td>
<td>4</td>
</tr>
</tbody>
</table>

Figure 3. Number of Knowledge Organization authors per year by country. Full visualization can be accessed online: https://public.tableau.com/profile/heather8449#!/vizhome/MappingKOauthorship/Authorsperyearbycountry.
Below the national level, we coded authors in terms of institutions, usually universities, and, where provided, with the academic unit such as school, department, college, etc. Taking these names supplied by authors, a broad overview of the disciplinary nature of home units can be generated. Although single instances of affiliations with departments of archaeology, for example, are not depicted in the word cloud generated, a sense of the most common departments is available from scanning Figure 6. “Information” is the overarching school/department name, with “library” and “computer” perhaps unsurprisingly next in proportion. Interestingly, “communication,” “management,” “engineering,” “economics,” “business,” and “technology” are also well represented, creating at least an initial sense that the view of KO as naturally interdisciplinary is supported.

Using Hofstede et al.’s (2005) dimension of individualism-collectivism (the spreadsheet of Hofstede dimensions used in this project was downloaded from the following source: https://geerthofstede.com/research-and-vsm/dimension-data-matrix/), authors publishing in Knowledge Organization from countries ranked on this dimension can be compared to the average number of authors on articles. In Figure 7, the darker the color of the country, the higher the “individualism” index score. As Hofstede et al. remark (90), “The vast majority of people in our world live in societies in which the interest of the group prevails over the interest of the individual,” but it is clear that significant national differences exist. Knowledge Organization has a great deal of interest from authors in what Hofstede et al. deem more “individualist” cultures, including Canada, the United States, Great Britain, and Australia. In fact, Australia, a highly individualistic country, averages one author for paper (N.B., only two papers with an author from Australia were included in the dataset). Farther along the spectrum of the individualism-collectivism dimension is China, a more collectivist culture in Hofstede’s survey, and indeed Chinese scholars publish papers with an average of over three authors.

When the average number of authors per article by country is plotted against a country’s individualism-collectivism dimension score, the trendline reinforces the idea that countries with a higher individualism score like Canada, Great Britain, the United States, and Australia (averaging between roughly one and two authors per article) have fewer average authors per article than more collectivist countries such as Colombia and Pakistan, which average four authors from their country per article. See Figure 8. The graph, however, is anything but neat, with the bulk of the articles having between one and three authors regardless of country of origin. The data in Figure 8 also represent variations introduced by other cultural dimensions, but nonetheless, even with the caveats we might place on the Hofstede model and the limited data set of Knowledge Organization authorship, these trends present an interesting lens on authorship and co-authorship.

Figure 4. Average number of authors per article by country for 2009-2018 inclusive. An interactive version of this map is available online: https://public.tableau.com/profile/heather8449#!/vizhome/MappingKOauthorship/Authorsbycountry.
Figure 5. Distribution of authorship by country for each year of study. An interactive version of these maps is available online: https://public.tableau.com/profile/heather8449#!/vizhome/MappingKOauthorship/Timelapse2009-2018.
Figure 6. Word cloud showing alphabetical list of the top fifty of 223 possible words from department or school names with stop words in a number of languages applied (generated using https://tagcrowd.com/).

Figure 7. Individualism and authorship by country. An interactive version of this map is available online: https://public.tableau.com/profile/heather8449#!/vizhome/MappingKOauthorship/HofstedeIndividualism.
Beyond the rate of single or co-authorship, we might ask if the interests of authors in individualistic and collectivist cultures are similar or different? Based on the author’s country of residence, deduplicated lists of the first lines of article titles were used to create word clouds for a group of collectivist countries with individualism dimension scores between 18-26, all of which are in East Asia (see Figure 9). A second word cloud was created based on titles of articles by authors based in the United States (see Figure 10). For both, the term “knowledge” was removed given its frequency in all papers. The East Asian titles represent a smaller set of words (113 possible words) and show greater cohesion, with more words displaying with larger font, indicating frequency of use across titles. The presence of “Chinese,” “Mekong,” and “national” suggest perhaps a concern for local initiatives. Interestingly, the term “organization” does not appear in the East Asian list, which is somewhat surprising given this journal’s coverage. In the US titles (a set of 287 possible words), “organization” is predominant, with “analysis,” “domain,” and “ethical” the next most common title terms. Again, one should not draw too firm a conclusion from these trends but they suggest some differences in emphasis on KO scholarship across regions and cultures.

Lastly, in considering the geography of contributions and relative wealth, Figure 11 presents a map where countries with larger GNIs are indicated in darker green. Is there a wealth threshold for Knowledge Organization authors? Is KO the province of richer or wealthier nations? Contributions seem to be somewhat balanced and there is a range of countries on the wealth index participating in KO but this is clearly a challenge in all disciplines and one that might be usefully explored further in terms of Knowledge Organization’s global growth and reach.

4.0 Conclusion

This research presents a first pass at characterizing the international and interdisciplinary community of scholars publishing in Knowledge Organization. This preliminary anal-
Figure 9. Word cloud showing alphabetical list of the top fifty deduplicated article title words, “knowledge” removed, from countries with individualism indexes 18-26 (i.e., Malaysia, China, Thailand, Singapore, and South Korea) (n=29) (generated using https://tagcrowd.com/).

Figure 10. Word cloud showing alphabetical list of the top fifty deduplicated article title words, “knowledge” removed, from the United States (individualism score ninety-one) (n=98) (generated using https://tagcrowd.com/).
ysis suggests four conclusions, with some caveats, as follows:

The publication base is growing. Over the last decade there has been a generally upward growth in the number of articles published in *Knowledge Organization*, with the article count doubling from 2009-2018.

KO research is now a global activity, with published papers coming not just from the established scholarly communities in Europe and North America but from China and other parts of Asia, the Middle East, South America, Africa, and Australia. While the numbers in some regions are low, there is reason to be optimistic that KO is establishing itself internationally as a discipline.

Authorship patterns indicate that co- or group-authorship is routine, but the trend in these numbers suggests the broad individualist-collectivist distinction of cultures by Hofstede might help us understand the primary differences among regions on this variable.

Topical analysis suggests that research in KO may also reflect global cultural differences, particularly on the individualist-collectivist dimension of Hofstede et al. Our data focused only on two particular regions but is not exhaustive.

There are clearly several limitations to this work. First, we are using data from only one journal. KO is a field practiced outside of English-speaking areas and thus the contributions of non-English language scholars are invisible to this project. Further, this is but a preliminary analysis, using a limited number of measures for a reduced data set of only ten years. While we intend to complete the analysis on the full set of back issues, fewer research papers were published in the early years. Ideally, we would like to compare KO with other areas within information science to determine if *Knowledge Organization* is unique in its pattern of authorship and global activity. Finally, while broad examination of author patterns is interesting, it would be instructive to add a deeper thematic analysis to identify trends in coverage or topics that might indicate how *Knowledge Organization* is evolving over time as well as across regions. It is important to recognize also that direct conversations with authors, particularly those from different regions, would complement this analysis in terms of author motivations, perceived challenges, and sense of intellectual identity in KO. In sum, we believe there is more work ahead but the early indications are that such analyses of disciplinary records can prove insightful for information scientists.

References

Darmani, Anna, Nidal Dwaikat and Andres Ramirez Portilla. 2013. “Twelve Years of Scholarly Research: Con-
Episemantics:
Aboutness as Aroundness†

Elliott Hauser* and Joseph T. Tennis**

*University of North Carolina at Chapel Hill, School of Information and Library Science,
100 Manning Hall, Chapel Hill, NC 27599-3360, <eah13@email.unc.edu>
**University of Washington, The Information School, Box 352840,
Mary Gates Hall, Ste. 370, Seattle, WA 98195-2840, <jtennis@uw.edu>

Elliott Hauser, MSIS, is a doctoral candidate and at the University of North Carolina at Chapel Hill, where he is a member of the Royster Society of Fellows. His research interests include information organization, data science education, the ethics of algorithms, and sociotechnical methods of systems analysis. His dissertation research on the phenomenon of certainty in information systems was recently selected for the Litwin Books Dissertation Research Award. Elliott has a background in programming, co-founded the online coding education platform Trinket, and teaches programming and information studies topics at the graduate and undergraduate levels.

Joseph T. Tennis is Associate Professor and Associate Dean of Faculty Affairs at the University of Washington Information School, Adjunct Associate Professor in linguistics, and a member of the textual studies, computational linguistics, and museology faculty advisory groups at the University of Washington. He served as President of the International Society for Knowledge Organization from 2014-2018. He is on the Library Quarterly and Knowledge Organization editorial boards and served as a core member of the InterPARES Trust research team from 2005-2019. Tennis works in classification theory, metadata versioning, ethics of knowledge organization work, descriptive informatics, and authenticity.


Abstract: Aboutness ranks amongst our field’s greatest bugbears. What is a work about? How can this be known? This mirrors debates within the philosophy of language, where the concept of representation has similarly evaded satisfactory definition. This paper proposes that we abandon the strong sense of the word aboutness, which seems to promise some inherent relationship between work and subject, or, in philosophical terms, between word and world. Instead, we seek an etymological reset to the older sense of aboutness as “in the vicinity, nearby; in some place or various places nearby; all over a surface.” To distinguish this sense in the context of information studies, we introduce the term episemantics. The authors have each independently applied this term in slightly different contexts and scales (Hauser 2018a; Tennis 2016), and this article presents a unified definition of the term and guidelines for applying it at the scale of both words and works. The resulting weak concept of aboutness is pragmatic, in Star’s sense of a focus on consequences over antecedents, while reserving space for the critique and improvement of aboutness determinations within various contexts and research programs. The paper finishes with a discussion of the implication of the concept of episemantics and methodological possibilities it offers for knowledge organization research and practice. We draw inspiration from Melvil Dewey’s use of physical aroundness in his first classification system and ask how aroundness might be more effectively operationalized in digital environments.

Keywords: meaning, aboutness, classification, subject, episemantics

† Presented at NASKO 2019: Knowledge Organization: Community and Computation, sponsored by ISKO-Canada/United States at Drexel University, Philadelphia, Pennsylvania, USA, June 13-14, 2019. We are grateful to an anonymous reviewer for encouraging us to explore the literal aroundness central to early library classification schemes.

1.0 Introduction

This paper discusses and synthesizes two conceptions of the term episemantics developed independently by the authors in prior work. Both conceptions deny that meaning is an inherent property of language, but take distinct approaches in relating this idea to the field of KO, and information studies more broadly. Tennis (2016) proposes episemantics as a potential new field of study, analogous to epigenetics, just recently made possible due to the advent of new technologies and research methods. Hauser (2018a) asks what it might mean to remove aboutness as a core com-
ponent of our understanding of information at all. After discussing both proposals, we present a synthesis of each that connects Tennis's methodological proposal with Hauser's theoretical approach via a shared pragmatism, in Star's sense of “consequences, not antecedents.” The result is discussed in relation to classification theory and particularly in light of Melvil Dewey's pragmatic approach to his first classification system. Finally, we consider what this might mean for organization practices in digital environments.

2.0 Tennis's episemantics: Epigenetics for KO

The idea of episemantics is to account for meaning as it changes over time outside of the scheme, and relate that to the scheme. Instead of reifying the subject in the context of the scheme alone, and linking those subjects to a body of documents, episemantics would establish models for the investigation of particular relationships. These models would be networks of meaning that show how relationships between terms are established.

Tennis, “Methodological Challenges in Scheme Versioning and Subject Ontogeny Research,” 578

Tennis employs an analogy to epigenetics, the study of the effects and behavior of genetic material within living organisms, as opposed to limiting the scope of study to “a” genetic sequence. Epigenetic research has determined that the activation and inhibition of specific genes often occurs in response to environmental or organismal factors in what must be regarded as emergent properties not detectable from a mere sequence of nucleotides. Just as rapid and inexpensive sequencing techniques allowed the relative rates of expression of genes to be contemplated as a subject of research, thereby enabling a new field, Tennis envisions digital methods providing new epistemic access to phenomena of deep importance to subject ontogeny research.

The challenge that Tennis's proposed episemantics addresses is the “location” of meaning in indexing languages in relation to literary warrant. Most indexing languages rely on their structure and the intellect of the indexer to triangulate the meaning in indexing terms. Further, meaning can be inferred from the range of materials that are indexed with that term. What has heretofore been lacking is the link to the literature except in the rare cases of citations to literature in thesauri (Soergel 1974) and Library of Congress Subject Headings (e.g., Library of Congress 2019). However, there are no explicit links between these sparse citations and wider network of literature.

Elsewhere, Tennis has presented on the circumstantial evidence relating term appearance in the Dewey Decimal Classification to literary warrant using the Google Books and Hathi Trust corpora (Tennis 2012). Constructing an episemantic methodology would allow for explicit links, revealing how terms were deployed in literature.

Essentially, Tennis's exploratory proposal would allow subject ontogeny researchers to connect the meaning of subjects to both the use of those subject terms (via large scale analysis of cataloging records) and the separate use of the same terms outside the context of knowledge organization (via the methods of corpus linguistics). While these methods do not eliminate the methodological concerns Tennis has identified (2016), they represent viable new lines of research with implications for concepts of aboutness and meaning within the LIS context. This would be a nod to studying the semantics and the pragmatics (in the linguistic sense) of terms alongside their role in indexing and in warrant. Analysis of the “code” of indexing languages in KO could thus be substantially supplemented by examinations of its emergent “expression” within works and records at scale. We will elaborate on this possibility below.

3.0 Hauser's episemantics: posterior projection of meaning

Losee's conception of aboutness's role arises from a category error: while processes’ output is related to both their input and the processes themselves as he claims, that relationship should not be described as aboutness until episemantic interpretation occurs. Following logical empiricism, Losee assumes that episemantic interpretation is (or: can be; should be; for science, must be) a transparent process, enabling processes’ outputs to be about their inputs. I contend that aboutness only obtains in the relationship between the interpretation process and the jussive encoding process.

Hauser, “Information from Jussive Processes,” 303

Influenced by both the pragmatic philosophy of language and its continental critics, especially Derrida, Hauser emphasizes the lack of meaning inherent to inscriptions. For Hauser, this is encapsulated in Bowker's discussion of the jussive. Bowker views “memory practices” in light of the way in which they enact forgetting (Bowker 2006; Hauser 2018b). For Hauser, this amounts to a proposal to investigate technologies of remembering via the techniques of forgetting they enable.

These observations were sparked by a critique of Losee, who seeks to embed an informative aboutness into a domain-independent account of information (Losee 1997, 2012). Losee renders information as the result of pro-
cesses and as informative “about” the process and its inputs. This is a powerful approach but problematically embeds a strong representational aboutness within the foundation of information. While scientific realists are likely to see no problems with such an arrangement, Hauser seeks to preserve the power and expansive domain of Losee’s work while stripping it of its reliance on scientific realism. Scientific realism is incompatible with many of the domains we serve, so Hauser tries to preserve as a possible viewpoint while avoiding placing it at the core of our discipline.

Contra Losee, Hauser locates aboutness as subsequent to the interpretation of inscriptions rather than as inherent to processes. Episemantics is thus the posterior projection of meaning (and aboutness) onto inscriptions via interpretation. Meaning is always enacted rather than inherent. This includes both the meaning of information resources and of indexing languages. To revise Losee’s formulation, information is merely subsequent-to processes; aboutness comes afterwards according to Hauser (2018a, 304): “The aboutness relationship consists of and is created by the episemantics of interpretation.” Aboutness is thus not a property but a relation that arises out of interpretive acts.

Though it is inherently constructivist, Hauser takes pains to situate scientific realism within this conception. In Hauser’s reading, Losee’s information from processes and its embedded aboutness results from a specific account of the process of interpretation. “Following logical empiricism, Losee assumes that episemantic interpretation is (or: can be; should be; for science, must be) a transparent process, enabling processes’ outputs to be about their inputs” (Hauser 2018a, 303). This framing doesn’t exclude strong representationalist conceptions of aboutness but rather de-centers them. They are one amongst many potential instances of the creation of meaning. It is this de-centering which accomplishes Hauser’s pluralistic goal. As, for better or worse, a metadiscipline (Bates 1999), we must serve a variety of compliances Hauser’s pluralistic goal. As, for better or worse, the creation of meaning. It is this de-centering which ac-

4.0 Episemantics, recombined

Each conception could stand on its own, but we’ve found it generative to consider how the two conceptions might be recombined. Methodology and theory should ideally reinforce each other’s strengths to form a coherent whole. Can such a project be accomplished here?

Tennis’s account is much more deeply embedded within the methodology of classification research, especially subject ontology. This depth makes it clear how it might be applied, but obscures the true power and breadth of the idea. Hauser’s approach is more general. This generality offers greater breadth but is ultimately diffuse and difficult to apply. This section will show how the two approaches can be combined to maximize their strengths and mitigate each other’s weaknesses.

Tennis’s analogy to epigenetics is apt, and a closer look at the field of epigenetics offers an important template for how KO might evolve like traditional genetics when confronting these ideas. Traditional genetics might simplistically be thought of as a series of sophisticated rules for labeling organisms and groups of organisms. Medical genetics uses the possession of genes as, effectively, a categorization rule to inform statistical analyses of morbidity and mortality (e.g., patients with this gene are X% more likely to develop heart disease, and live, on average, Y years less than those without). Phylogenetics uses algorithmic measures of similarity to infer ancestral relationships between species. Each of these approaches contains a step when the object of study is simply labeled genetically, and from this point on the label is all that is available. This labeling process is jussive, in Bowker’s sense, and encodes a specific disciplinary technique of forgetting.

Epigenetics represents a deepened interpretation of DNA sequences by bringing their expression into view. Traditional genetics was presumed to be a method for finding the animating code behind everything but at times has devolved into a sophisticated mechanism for tagging data prior to statistical analyses of co-occurrence patterns. Epigenetics has a claim to this original promise, but must do so by abandoning a view of genetic sequences as determinants of the futures of the organisms that possess them in favor of a more fully contextualized account of how those genes proliferate and are expressed within an organismal and ecological context. Wendy Chun has noted the logocentrism common to biology and computing technologies (Chun 2013). She makes the novel, but convincing, claim that the kind of logodeterminism represented in works like Schrodinger’s “What is Life” was an important precursor to our understanding of what code is and how computers work (Chun 2013, Ch. 3). Chun’s analysis suggests a new light within which to view Tennis’s analogy: that episemantics might offer a path, parallel to that of
ics, whereby we gain a greater account of context, and greater explanatory power, by abandoning an outdated logodeterminism.

Such a potential approach is offered by a pragmatic account of aboutness as aroundness, or “in the vicinity, nearby.” The occurrence of a gene within a strand of DNA is irrelevant unless the gene is expressed. The import of a gene remains unknown precisely “until” we have an account of its expression within an organism, population, or ecological context. Genetic expression is a process of interpretation within a context. Similarly, the possession of a term within an indexing language, or applied to a specific work, is irrelevant until we know how such a term is used. The meaning of a term is impossible to analyze prior to a contextualized account of use.

Thus, pragmatism forms a bridge between Tennis and Hauser’s accounts of episemantics. Pragmatism implicitly animates a good deal of LIS work, and has recently gained traction as a subject of research in its own right (Douša 2009; Buschman 2017; Sundin and Johannisson 2005). While competing accounts of pragmatism have been offered, we prefer Star’s simple and concise definition: a focus on “consequences, not antecedents” (Star 2015, 133). Star here references the words of her mentor Anselm Strauss, who in turn was inspired by the work of John Dewey. This pragmatic ethos unites all three thinkers, even as the meaning of this mantra has evolved. Bowker and Star’s Sorting Things Out would have been far less impactful for our field if it had been subtitled Classification and its Antecedents. A focus on consequences animates both Hauser and Tennis’s approaches. Tennis’s epigenetics analogy shifts focus away from the antecedent, DNA-like indexing language to the consequent, RNA-like classification records and the content of the works they classify. Hauser positions the antecedent inputs of an informative process as ultimately irrelevant to the aboutness of the consequent interpretation and enactment of output.

5.0 Why does KO need an account of episemantics?

Episemantics represents an important reminder to avoid viewing meaning as an inherent property of either indexing terms or abstract concepts. This offers the key methodological benefit of a shared account of both natural and artificial languages in a way that concepts like literary warrant cannot. The materiality of language emphasized by Hauser acts to blur the distinction between natural language, indexing languages, and computer languages. This, combined with Tennis’s proposal to look for traces of use within all three kinds of languages, presents a new picture of what classification research might become. In addition to strengthening existing techniques such as subject ontogeny, our recombined concept of episemantics offers a glimpse of what larger scale, comparative “subject phylogeny” might be.

If we take episemantics seriously, we must revise our conception of aboutness. The notion of meaning somehow inhering in the inscriptions that constitute a language (what Star might call an “antecedent” view of meaning) has proven philosophically problematic for human languages. Given this difficulty, we suggest abandoning an attempt to clarify or utilize this traditional sense of aboutness for indexing and computer languages. Instead, a turn to pragmatism about meaning and a focus on investigating use, both within narrow contexts and at scale, offers a viable way forward. “Aboutness” in this view need play no larger role than suggesting that something has been placed near something else, as librarians commonly do with cataloged books. The effects of cataloging may be deeply complex, socially embedded, and ethically significant, but the analysis need not include a strong account of aboutness as inherent meaning. Rather, we argue, an episemantic approach precludes this.

Our proposal does not seek to or need to enforce a uniform account of aboutness to succeed. Researchers who still believe that a strong account of inherent meaning is possible may continue to pursue work in that direction separately. To move forward, we need only agree to proceed with a weak aboutness within the empirically and historically oriented study of classification. When we do, Tennis’s proposal of exactly how this might be studied at scale, for both subject ontogeny and the as-yet-unrealized field subject phylogeny, becomes merely a promising suggestion of many potential ways forward.

6.0 Aroundness, Dewey, and the digital

Although his classification system is often conflated with universalist classification projects, Melvil Dewey himself never considered the “aboutness” of his original classification system to be a specification of the property of the works cataloged and arranged on shelves. In the preface to the first edition of his classification, it is clear that his focus was primarily on the “effects of placing books near each other” (Dewey [1876] 1976):

In all the work, philosophical theory and accuracy have been made to yield to practical usefulness. The impossibility of making a satisfactory classification of all knowledge as preserved in books, has been appreciated from the first, and nothing of the kind attempted. Theoretical harmony and exactness has been repeatedly sacrificed to the practical requirements of the library or to the convenience of the department in the college.
The effects Dewey considered, of course, were both upon patrons when browsing the shelves, and upon the operation and maintenance of the library itself. The fact that subsequent versions of his system and its presentation became indelibly associated with universalist classification schemes need not prevent us from returning to it for inspiration. Dewey’s system, embedded as it was within the late 19th century library movement’s goals and cultural assumptions (for more on this, see Miksa 1998), was nevertheless a novel and pragmatic take on how to organize a newly abundant information resource for optimal use and management. Viewed in this lens, Dewey’s principle was to identify a physical property of information resources, their physical location, and produce a system for manipulating this property to balance the needs of library patrons and library staff. Though this system contained subject headings, these were merely cogs in an ultimately spatial machine. In our terms, this machine manipulated aroundness rather than ascribing aboutness.

Recapitulating this approach with digital resources is non-trivial. Unlike a physical library, the interfaces, sequences, and formats that users access digital information are wildly disparate. To give a simple example, library patrons walk through the front door. Taking this into account, libraries could arrange resources in such a way as to reliably shape these first interactions. Though digital libraries still have putative “home” pages, users may land upon practically any part of the site, from practically any other digital context. What can serve the function that physical proximity did in Dewey’s original system?

This, of course, is a question with proliferating answers. In a sense, the intractability of organizing the massive amounts of highly specialized knowledge, a task increasingly confronted by Dewey’s successors, encourages the essentialist approach to “aboutness” that we have critiqued. For a specialist researcher seeking journal articles in her speciality, a given resource is either “about” “the desulfurization of hot coal gas with regenerable metal” or not. As Miksa notes, classification theorists who took up the devilish challenge of organizing specialist knowledge, such as Richardson, Bliss, and Rangagathan, found themselves increasingly drawn to map a “universe of knowledge,” where every specialist query could have a definite home (Miksa 1998, 56–73 et seq.). Through the lens with which we have been reading Dewey’s work, this strikes us as precisely an attempt to provide an analogy to the physical location that made Dewey’s system work for generalist libraries. A conceptual location within the Cartesian space of the universe of knowledge would, modernist classification theory held, allow the precise provision of the right resource for any sufficiently specified need.

The task of repeating this process without universalization and its attendant definite aboutness is one we suggest as a future research program. Methodologically, Tennis’s proposal of utilizing large scale computational linguistics as a kind of window into the use and relationship of words to each other in a corpus would help ground such a project in the actual use of language rather than encouraging the invention and perfection of a crystalline representation of the universe of knowledge. Hauser’s exhortation to remove meaning from classification helps us uncover the practical effects of classification activities. Dewey’s pragmatism led him to focus on the physical arrangement of books. Subsequent modernists sought an ideal, universal space within which to arrange and relate classes to each other. The fragmented space of new digital technologies belies either approach. Knowledge is not a set of cartesian coordinates, waiting to be arrayed in crystalline perfection. There is no reliable experience of physical space to structure patrons’ encounter with digital resources. How might we re-envision these organization practices to instead modulate properties that acknowledge the fractured nature of digital encounters but provide flexible structure for navigation and exploitation of digital resources?

7.0 Conclusion

In two separate threads, Tennis and Hauser point to a contingent and pragmatic view of aboutness. This leads us to reconsider the concept in terms of an earlier meaning, “in the vicinity, nearby; in some place or various places nearby; all over a surface.” The vicinity and surface of meaning, we have argued, are epistemantically derived—both theoretically and methodologically.

Revisiting the early work of Dewey, we uncovered a new sense of aroundness, a literal one. Physical location was central to Dewey’s scheme to balance the needs of patrons and library staff. Modernist classification theorists, who Miksa read as constructing “the universe of knowledge” as their domain, still employed an attenuated aroundness in their schemes relating classes, and thereby subsequently cataloged resources, to each other via their physical proximity within collections. Dewey’s pragmatism centered around the realization that physical location was the primary “outcome” of his classification and the primary tool he had to influence library operations.

In a digital environment, many possible operationalizations of aroundness are possible. Commercial information systems have pioneered many of these, driven by large scale collection of user data (“Customers who viewed this also viewed”). The synthesized conception of episemantics advanced in this paper is intended to support deep engagements with these new possibilities. We hope that a pragmatic analysis of the consequences of different manifestations of aroundness might help provide guidance for continued innovation in KO.
And, of course, episemantics remains an exciting methodological proposal for subject ontogeny research. Hauser's theoretical contributions are consonant with the goals of subject ontogeny. If meaning is viewed as enacted interpretation, examining traces of such interpretation in corpora and even individual cataloging decisions helps provide insights unavailable by any other means. Amongst its other goals, we hope that this paper encourages the large-scale collaborations needed to understand the complexity of semantic construction that animates KO activities, now and in the past.

References


The Respective Roles of Intellectual Creativity and Automation in Representing Diversity: Human and Machine Generated Bias †

Vanda Broughton

University College London, Department of Information Studies, Gower Street, London WC1E 6BT, <v.broughton@ucl.ac.uk>

Vanda Broughton is Emeritus Professor of library and information studies at University College London. Her principal research interest is in the development of faceted classification, particularly as it affects different disciplines. She is editor of the second edition of the Bliss Bibliographic Classification (BC2), and an associate editor of the Universal Decimal Classification. In addition to the published volumes of BC2, she is author of several books on knowledge organization systems and numerous articles and conference papers.


Abstract: The paper traces the development of the discussion around ethical issues in artificial intelligence, and considers the way in which humans have affected the knowledge bases used in machine learning. The phenomenon of bias or discrimination in machine ethics is seen as inherited from humans, either through the use of biased data or through the semantics inherent in intellectually-built tools sourced by intelligent agents. The kind of biases observed in AI are compared with those identified in the field of knowledge organization, using religious adherents as an example of a community potentially marginalized by bias. A practical demonstration is given of apparent religious prejudice inherited from source material in a large database deployed widely in computational linguistics and automatic indexing. Methods to address the problem of bias are discussed, including the modelling of the moral process on neuroscientific understanding of brain function. The question is posed whether it is possible to model religious belief in a similar way, so that robots of the future may have both an ethical and a religious sense and themselves address the problem of prejudice.

Received: 18 September 2019; Revised: 12 November 2019; Accepted 14 November 2019

Keywords: machine intelligence, bias, human, artificial intelligence, data


1.0 What is artificial intelligence?

There are many and varied definitions of artificial intelligence, and various synonyms for it. Poole et al. (1998, 1) note that “the term ‘artificial intelligence’ is a source of much confusion,” preferring to call it “computational intelligence,” although it is likely that artificial intelligence is today the more widely recognised term. Other names include “machine intelligence,” “synthetic intelligence” (Brachmann 2005; Gorg et al. 2014), and “augmented intelligence” (Ojala 2018; Albrecht et al. 2015; Hannay 2014).

The Encyclopedia Britannica (Copeland 2019) defines artificial intelligence in the following manner:

Artificial intelligence (AI) [is] the ability of a digital computer or computer-controlled robot to perform tasks commonly associated with intelligent beings.

The term is frequently applied to the project of developing systems endowed with the intellectual processes characteristic of humans, such as the ability to reason, discover meaning, generalize, or learn from past experience.

The same article identifies five key aspects of intelligence, whether human or machine: learning, reasoning, problem solving, perception, and language. At the operational level, machine intelligence can take a number of forms: pattern recognition, voice recognition, image (including facial image) recognition, and machine translation. Copeland’s definition (above) tends towards the narrower field of machine learning: “a form of AI that enables a system to learn from data rather than through explicit programming … “After a model has been trained, it can be used in real time to learn from data” (Hurwitz and Kirsch 2018, 4-5).
For the purposes of this paper, artificial intelligence is considered mainly within the context of information retrieval, specifically document retrieval, and the ways in which document content can be automatically identified using intelligent agents. This may involve the construction of automatic classifiers through machine learning and the way in which document content is processed.

2.0 Artificial intelligence as a complement to human activity

Artificial intelligence has impacted on many areas of human activity, in part because of the speed with which it can process information in an overloaded world, saving human effort and apparently offering a more objective way to assess and respond to a variety of situations.

Information management is only one of such uses of AI, where it may promise to solve the perennial problem of organization and retrieval in a situation where there is “too much to know.” This situation has been acknowledged since the early modern period (Blair 2010), began to be addressed by mechanization in the mid-twentieth century, and in the twenty-first century prompted numerous studies of the way in which machines might automatically analyse, categorize, index, and classify documents and other information objects through a process generally referred to as automatic metadata generation or AMG (Broughton, Palfreyman, and Wilson 2008; Greenberg et al. 2005).

During that period there had also been a good deal of research into the relative roles of controlled vocabularies and automatic indexing, generally leading to the conclusion that a hybrid model offered the best balance between efficiency and effectiveness; a number of studies demonstrated that the use of a controlled vocabulary improves the performance of the tool or system (Liang et al. 2006; Cheung et al. 2005; Aula and Kaki 2005; Ko et al. 2004). Another major theme was the automatic building of classificatory structures such as ontologies, independently of humans, from text corpora or other sources. The extraction of data from text continues to be a common means of constructing semantic tools, but, as we may see below, the assumption that terms in text are value free, and mean exactly what they say, presents a danger to the usefulness and efficiency of such exercises.

As AI gains ground as an established tool for processing and decision making, particularly with respect to personal data, some questions have been raised as to the acceptability of AI in this role, the ethics of AI, and the extent to which machines can function as intelligent, and as ethical agents. Associated ideas, such as the personhood of robots, and whether they can be said to assume responsibility for their actions, have also been considered.

3.0 Ethical considerations in knowledge organization

It is now well established that the business of knowledge organization, whether that is classification, indexing, subject representation through headings, or visualization tools, brings with it some ethical concerns. Recently, attention has focussed on fake news, or controversial thinking, such as holocaust denial, and how such material should be represented, but more generally concerns are with the misrepresentation or under representation of minority groups, leading to disadvantage and disempowerment. There is now a growing body of literature on the ethical theory and philosophy of KO (Olson 1998; Szostak 2014; Mai 2010, 2013a, 2013b, 2016), and a substantial number of studies of the way in which it can discriminate on the basis of gender (Foskett 1971; Marshall 1977; Olson and Ward 1997; Olson 2007), sexual orientation (Drabinsky 2013; Fox 2016; Howard and Knowlton 2018), race and ethnicity (Duarte and Belarde-Lewis 2015; Adler and Harper 2018), political status (Lacey 2018), and religion (Broughton 2000; Broughton and Lomas 2019).

All of such bias is problematic in a world of increasing diversity, and the major players in conventional KO are seen to address some of the worst excesses. Factors which exacerbate the bias include: unequal provision either of terminology or (in a coded system such as a classification) unequal distribution of notation; failure to name at all certain groups or perspectives; and language which has a strong flavour of one particular favoured perspective or culture. Where culture is a powerful element, as in religions, language is a specific problem.

4.0 Ethical considerations in artificial intelligence

There has been substantial research into the phenomenon of machine ethics, that is the potential ethical or moral behaviour of intelligent agents. It should be carefully differentiated from computer ethics which is concerned with the behaviour of humans in the context of computing and information technology, and with roboethics which refers to ethical behaviour of humans in the design and construction of intelligent machines, and in human-machine interaction. Although there are some twentieth-century discussions of the possibility of moral—or immoral—actions of machines, the field really begins with the 2005 AAAI Symposium on Machine Ethics, where the problem is clearly stated, and named by Anderson et al. (2005):

Past research concerning the relationship between technology and ethics has largely focused on responsible and irresponsible use of technology by human beings, with a few people being interested in how
human beings ought to treat machines. In all cases, only human beings have engaged in ethical reasoning. We believe that the time has come for adding an ethical dimension to at least some machines. Recognition of the ethical ramifications of behavior involving machines as well as recent and potential developments in machine autonomy necessitate this. We explore this dimension through investigation of what has been called machine ethics.

In Anderson et al.’s paper, they consider the implementation of two systems of machine ethics, based on philosophical principles as displayed in the work of W. D. Ross (theory of prima facie duties), and Jeremy Bentham (Utilitarianism), both of which can be expressed as a series of rules. Utilitarian ethics and the basis of its decision-making is of particular interest, since in the form of Bentham’s *Felicific calculus*, or *Calculus of pleasures* (1789), it was designed to be computable, and indeed, one of the themes of the Symposium was the computability of ethics. At the time of Anderson et al.’s research, the likely guarantee of “good” machine ethics was the imposition of better and more considered rules for the machine’s operation, derived from traditional systems of ethics and the practice of professional ethicists. As they say in a subsequent paper (2007, 25):

Ensuring that a machine with an ethical component can function autonomously in the world remains a challenge to researchers in artificial intelligence who must further investigate the representation and determination of ethical principles, the incorporation of these ethical principles into a system’s decision procedure, ethical decision making with incomplete and uncertain knowledge, the explanation for decisions made using ethical principles, and the evaluation of systems that act based upon ethical principles.

### 4.1 Where machine ethics falls short: bias in intelligent agents

The general assessment of machine information processing and machine decision-making has been that it may avoid the subjectivity associated with humans. In practice this has turned out not to be the case, since, despite the emphasis on machine independence in artificial intelligence, intelligent agents are not created spontaneously, but require some degree of human participation, and no system of machine learning can avoid the use of information which has been at some stage processed by humans. The problem affects equally machine learning where the agent has learned from data or a prepared model or training set, or in the case of knowledge organization systems, where human-constructed vocabularies or ontologies have been sourced by the agent. Additionally, human intervention often supports the machine-learning process through iteration with the “teacher,” usually through a technique of query-by-example accompanied by feedback to the machine.

Too often the result of this human input is that the machine inherits the prejudices of the human, so that the bias is hard-wired to the machine (Crawford 2016; Kirchner et al. 2016; Sears 2018; Kochi 2018). The World Economic Forum (2018, 3) stated:

> Designed and used well, machine learning systems can help to eliminate the kind of human bias in decision-making that society has been working hard to stamp out. However, it is also possible for machine learning systems to reinforce systemic bias and discrimination and prevent dignity assurance.

The likelihood of such bias has considerable implications for human rights, for the proper management of social diversity, and for the fair treatment of diverse groups in society. Such inequity is a long-standing problem in conventional information management and has been addressed at length in the research literature of knowledge organization in particular. It seems, however, especially insidious in the machine intelligence context, perhaps because of the expectation that higher levels of neutrality and objectivity apply.

### 4.2 Bias and discrimination derived from data

Much of the literature in this area is centred on machine decision-making based on demographic data, and the concern arises from a human rights perspective where some groups are disadvantaged or marginalized by the way in which the data is set up (Smith, Patil and Muñoz 2016; Obama White House 2016; World Economic Forum 2018). Generally in these cases, the data is factual and the decision-making based on demographic data, and the consideration and recognition of patterns embedded in data, especially latent associations between one group of attributes and another.

Particularly prominent in the discussion of bias in AI is gender discrimination, also a feature of early research into bias in KO. A recent major study by Criado-Perez (2019) reveals that data itself is often biased, because the sample is in some way flawed. Criado-Perez’s principal concern is with gender imbalance, and it is clear that a female perspective is often omitted, because the data is derived from studies that dealt only with males. Criado-Perez provides examples of where, for example, diagnostic thresholds based on biomarkers are inaccurate for women, because average figures are based on predominantly male data (as in Khamis et al. 2016), since the inclusion of female data
4.3 Bias and discrimination derived from semantics

A much-cited paper by Caliskan et al. (2017) establishes that not only is any incompleteness or skew in the data sample passed on to intelligent systems, but that semantics is also an inheritable factor. Using measurable associations between pairs of words, Caliskan builds on the work of some prior studies investigating human-like biases in textual corpora, particularly that of Greenwald (1998), who studied “biases that they consider nearly universal in humans and about which there is no social concern” (Caliskan, 183). Clear associations between “flowers” and “pleasant,” and “insects” and “unpleasant,” were replicated by Caliskan’s team, as were similar links between “weapon” (unpleasant) and “musical instrument” (pleasant), proving the soundness of the methodology. Caliskan et al. were also able to replicate more socially significant connections between European American names and African American names with pleasantness and unpleasantness respectively, a phenomenon confirmed in practice by Bertrand and Mullainathan (2004) who tested employers’ response to job applications varying only in the attached European or African sounding names.

Comparable work was also replicated in the area of gender, associating female names with “family” as opposed to “career,” when compared with male names (Nosek et al. 2002a), and the correlation of women with the arts, rather than mathematics, or with the sciences (Nosek et al. 2002b).

Studies of inherent discrimination based on religion are much less frequent, perhaps because religious affiliation is much less immediately obvious than gender or ethnicity. However, Binns (2018, 1) places it on a level with gender and race as a potential factor for discrimination, using the example of disparate treatment of nationals from Muslim-majority countries because of a perceived association of Islam with terrorism (2018, 4). Since such examples of religious prejudice are not uncommon and the problem is one with high public awareness, it is surprising that, to date, there is little or no research into bias associated with religious affiliation.

The existence of such semantic bias has considerable implications for information retrieval because so many automatic classifiers build structures on the back of text corpora on the assumption that these present a neutral and objective picture of the world. The existence of these biases seem to be clearly acknowledged in the world of corpus linguistics, but did not seem to be taken into account at all in the field of automatic classification or term extraction tools, perhaps because work on these originated in the sciences rather than the social sciences and humanities. Automated lexicography is now a very well-established methodology for building such semantic tools, whether the lexical data is sourced from other lexical tools such as dictionaries or extracted from text corpora, and the problem of inherited bias could consequently be a serious impediment to both effective retrieval and ethical practice.

5.0 Inherited semantic bias in religious terminology: the example of WordNet

WordNet is a vocabulary database, maintained at Princeton University, and used extensively as semantic content for all kinds of automatic indexing and classification tools. It defines itself in the following way (Princeton University 2010):

WordNet® is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. The resulting network of meaningfully related words and concepts can be navigated with the browser … WordNet’s structure makes it a useful tool for computational linguistics and natural language processing.

WordNet is displayed in a thesaurus-like format, although the tags it uses are not the conventional ones of information science, but rather offer a more analytical and nuanced range of inter-term relations. Nevertheless it approximates to the standard tags through its use of the categories hypernym (= broader term, superordinate class), hyponym (= narrower term, subordinate class), and synsets (equivalence relationships). Other relationships include types (= narrower term generic), instances (= narrower terminstantive), and meronymy (= narrower term partitive). Opening up a “sister term” reveals terms in the same array, comparable with some kinds of related, or associative terms. This suite of relationships provides the vocabulary with a robust logical structure, and verbs as well as nouns are thus organized into hierarchies. Unlike most controlled vocabularies, WordNet also includes adjectives and adverbs in its database. However there is quite limited use of associative term type links, so the navigation tends to be, on the whole, hierarchical.

WordNet is a good example of a resource that has inherited content. Although initially it was intellectually constructed, it draws on older sources such as thesauri (Barocas et al. 2018) that themselves may contain bias, and because of its widespread use in computational linguistics it passes on that bias.
In some applications, researchers repurpose an existing scheme of classification to define the target variable rather than creating one from scratch. For example, an object recognition system can be created by training a classifier on ImageNet, a database of images organized in a hierarchy of concepts. ImageNet’s hierarchy comes from Wordnet, a database of words, categories, and the relationships among them. WordNet’s authors in turn imported the word lists from a number of older sources, such as thesauri. As a result, WordNet (and ImageNet) categories contain numerous outmoded words and associations, such as occupations that no longer exist and stereotyped gender associations.

Even a cursory examination of WordNet’s religious categories reveals some very evident examples of bias. As with many humanities and social science disciplines, particularly those where there is a strong cultural dimension, language is a source of some problematic classes and linguistic expressions. Similarly, the precise analytical structure of WordNet, based on linguistics principles, while it is highly suitable for the sciences, does not always serve the rather messier humanistic domains nearly as well. Accurate and comprehensive category structure is not necessarily to be found, and in many cases the arrays are incomplete.

If we consider the standard criticisms of biased religion classes in standard bibliographic classifications, many of the same shortcomings are evident in WordNet. For example, if we look at the hierarchical display of hyponyms (subordinate classes) under Religion (disregarding the annotations and further levels of hierarchy) we find:

![Figure 1. Entry for “religion” in WordNet.](Image)
When compared with the standard “big twelve” religions acknowledged by most sources (for example, Hinnebels 2017; Boyett 2016), WordNet fails to mention Baha’i, Confucianism, Jainism, Sikhism (other than through its subset Khalsa), Zoroastrianism, and, amazingly, Islam. Expanding the list to include “full hyponyms” expands the hierarchy and brings in various Christian denominations, movements within Judaism and Buddhism, and under sects, Anglican High Church, Sunni and Shi’a Islam, the Society of Friends or Quakers, Jainism and Hare Krishna. Perhaps surprisingly, Scientology appears, but not the Mormon Church.

Needless to say, the omissions, odd associations and peculiar language (Hindooism looks very antiquated and mildly offensive) would be unacceptable in a modern thesaurus or bibliographic classification. Many of the definitions and verbal qualifications of entries exhibit some odd if not doubtful attitudes, as in the definition: “Hindooism (a body of religious and philosophical beliefs and cultural practices native to India and based on a caste system.” There are differing schools of thought about the caste system, and whether it arises from socioeconomic rather than religious forces, and this is a very contentious statement. Similarly, Paganism (synonyms: pagan religion, heathenism) is defined as “any of various religions other than Christianity or Judaism or Islamism” which is certainly inaccurate in respect of modern pagans, and potentially offensive to followers of the non-monotheistic faiths. Perhaps the worse sufferer is Islam, which is provided with the synonyms Muslimism, Mohammedanism, Muhammadanism, and Islamism. The Oxford English Dictionary says of Mohammedanism that “its use is now widely seen as depreciatory or offensive,” and none of these terms feel very appropriate or polite. It seems likely that this rather uncomfortable content has been imported from much older dictionaries without review or amendment.

Along with such archaic uses of language, there is also a leaning towards a strongly Christian-flavoured understanding of religious terminology, as opposed to a more multi-faith approach; examples of this can be seen in the table below. The terms have been chosen as relatively neutral ones which occur in a variety of religions, but in defining the terms or providing synonyms WordNet imposes a broadly Christian interpretation (see figure 2).

In fairness to WordNet, it does contain many religion specific terms (bhakti, Gemara, hajj, lama, menorah, nirvana, shaman, synagouge, etc.), but because of the mainly hierarchical structure these are not easily accessed through the parent religion. This positive feature needs also to be set against the general Christian tenor of the vocabulary and the evident tendency to view religion through a Christian lens.

There are also some straightforward factual inaccuracies in WordNet, for instance the qualifier of Hinduism “(the religion of most people in India, Bangladesh, Sri Lanka, and Nepal),” whereas the dominant religion in Bangladesh is Sunni Islam, followed by 83.4% of the population (Sawe 2019).

These significant shortcomings demonstrate a very considerable bias and a disregard for fairness and sensitivity towards minority groups. The gravity of bias in WordNet is considerably magnified by its widespread use as a lexical source for automatic classifiers, which implies that the prejudices will indeed have been inherited and reinforced by a great number of other intelligent agents.

<table>
<thead>
<tr>
<th>Source term</th>
<th>Synonyms</th>
</tr>
</thead>
<tbody>
<tr>
<td>altar</td>
<td>Communion table, Lord’s table</td>
</tr>
<tr>
<td>baptism</td>
<td>a Christian sacrament signifying spiritual cleansing</td>
</tr>
<tr>
<td>bless</td>
<td>make the sign of the Cross over someone</td>
</tr>
<tr>
<td>festival</td>
<td>religious festival, church festival</td>
</tr>
<tr>
<td>monk</td>
<td>Brother, Carthusian, Trappist, Cistercian</td>
</tr>
<tr>
<td>preaching</td>
<td>an address of a religious nature usually delivered during a church service</td>
</tr>
<tr>
<td>scripture</td>
<td>Bible, Christian Bible, Holy Writ, Word (the sacred writings of the Christian religions)</td>
</tr>
<tr>
<td>service</td>
<td>church service, prayer meeting, chapel service, vesper</td>
</tr>
<tr>
<td>sin</td>
<td>mark of Cain</td>
</tr>
</tbody>
</table>

Figure 2. Synonyms in WordNet.
As is the case in library and information science, where the interests and priorities of the user community demand a privileging of those interests, bias is not always regarded as a bad thing. It is generally agreed that the very fact of a specific perspective unwittingly and unavoidably generates bias towards the favoured group (such as classification schemes for libraries with specific religious affiliations). Given the importance of meeting user expectations and the needs of the user community, bias can be seen as an ethically-neutral phenomenon.

In other cases the investigation of bias is simply a part of the scholarly study of society and the legitimate search for patterns and trends in human cultures. For example, a paper by Kozlowski et al. (2019, 38) shows how the machine analytical technique of word embedding can help to reveal changes in social attitudes over time and historic changes in word meanings. In different situations, the identification of bias may be the preliminary to addressing it in a social and political context, and is a useful tool in highlighting social inequalities.

In a wider context however, bias should be energetically tackled if the system is not to appear as the tool of a particular cultural, political, or disciplinary community. Bias inherent in data is generally regarded as undesirable and has generated an area of research activity under the general heading of machine-learning fairness. Barocas, Hardt, and Narayanan (2018) provide a broadly-based survey of a number of problems and potential solutions, based on statistical adjustment. The book “offers a critical take on current practice of machine learning as well as proposed technical fixes for achieving fairness.”

Mancuhan and Clifton (2014) also propose a statistical solution to bias in data used for automatic financial decision-making, employing Bayesian techniques to identify and automatically correct bias. This is incidentally one of the few papers to reference religion as an attribute subject to bias, although the authors do not go on to include it in their study.

6.1 A moral and religious solution

As with every other area of human life, machine intelligence has impacted religious communities, apart from the general philosophical questions of whether robots can act as moral agents. A number of applications exist which aim to support religious practice, such as the Roman Catholic Confession app (Rau 2011) and Muslim Pro which can tell you prayer times and the direction of Mecca in your own town or village (Muslim Pro 2019), and attempts have already been made to use robots in ritual. Most of the literature here is in popular journals and the press, so it may be difficult to assess how serious these efforts are. We learn of a Christian robot priest in Wittenberg which radiates light from its hands and pronounces blessings in five languages as part of an exhibition to celebrate 500 years since the invention of printing technology, instrumental in the Reformation and the rise of Protestantism (Sherwood 2017). Other cases include a robot Buddhist monk in China (Tatlow 2016) which reads scripture and can answer questions, and another in Japan (Field 2017) which can “chant prayers and tap drums as part of a funeral ceremony.”

There is also a literature in the overlap between religious philosophy and AI that considers the nature of the relationships between intelligent agents, humans and the person of God, typically whether the creation of intelligent agents in some sense mirrors the creation of humans (Herzfeld 2003), and if the possibilities of transhumanism through the technological alteration of species are realizable (Dumsday 2017). Vidal (2007, 930) makes a comparison between man’s interaction with artificial beings and his interactions with the gods, asking whether the similarities are not caused by uncertainty:

But it is also true that where interaction is supposed to exist between the gods and their worshippers, there always remains a strong element of uncertainty which cannot easily be dismissed concerning the exact ontological nature of the hybrid arrangement by which the divinity’s presence is made manifest. It is precisely the same sort of ontological uncertainty that one finds expressed in the field of robotics. And this is also why robots both fascinate and worry the general public.

6.2 The moral and religious life of machines

A pressing question is whether a real sense of moral responsibility can be developed in intelligent agents, or, more fancifully perhaps, a proper religious sense. In human beings, it may seem obvious that ethical decisions differ in some significant respect from other kinds of decisions, and that intellectual reasoning is subordinate to, or at least strongly influenced by, emotional intelligence. As Liao (2016) says:

Central area of intellectual inquiry across different disciplines involves understanding the nature, practice, and reliability of moral judgments. For instance, an issue of perennial interest concerns what moral judgments are and how moral judgments differ from nonmoral judgments. Moral judgments such as “Torture is wrong” seem different from nonmoral judgments such as “Water is wet.” But how do moral judgments differ from nonmoral, but normative judg-
ments such as “The time on the clock is wrong” or “Talking with one’s mouth full is wrong”?

However, work in neuroscience has questioned whether this distinction between cognitive and emotional aspects of moral judgements is valid, suggesting instead that all such decisions depend on reasoning through complex calculation rather than a response to stimulus (Woodward 2016). Quartz (2009, 214) states:

> … recent findings suggest that the encoding of value in midbrain dopamine areas might underlie an early implicit encoding that is signaled to orbitofrontal cortex, where it guides choice.

Recent studies have shown that it is possible to identify precisely areas of the human brain responsible for social and moral behaviour, and to link deficits in brain function to immoral, or amoral, behaviour (Damasio 1994; Shoemaker 2012). Shoemaker (2012, 807) states:

> The basic limbic emotions are those present in all mammals emanating from phylogenetically analogous brain structures collectively called the limbic system … These are fear, anger, disgust, sadness, and happiness; they function chiefly to promote the survival of the individual. The moral emotions, the product of the social brain network, arise later in development and evolution (Adolphs 2003). They are guilt, shame, embarrassment, jealousy, pride, and altruism; they function to regulate social behaviors, often in the long-term interest of a social group rather than the short-term interest of the individual person (Adolphs 2003).

Such work has considerable implications for the development of ethically responsible machines, since if the nature of decision-making can be made explicit and the process modelled, then it is, at least theoretically, possible to replicate this process in machine decision-making.

A further and more difficult question is whether it is possible to inculcate religious sensibility in machines by a similar methodology. In the speculative Age of Spiritual Machines (1999, 6) Kurzweil, technologist and futurist, suggests that in the distant future this will happen spontaneously as the result of technological evolution:

> ... perhaps the most surprising finding to date is that core emotional structures, including the midbrain dopamine system and insula, decompose uncertain choice contexts along the statistical dimensions that are the cornerstone of FDT [Financial decision theory] … recent findings suggest that the encoding of value in midbrain dopamine areas might underlie an early implicit encoding that is signaled to orbitofrontal cortex, where it guides choice.

Even if we limit our discussion to computers that are not directly derived from a particular human brain, they will increasingly appear to have their own personalities, evidencing reactions that we can only label as emotions and articulating their own goals and purposes. They will appear to have their own free will. They will claim to have spiritual experiences. And people—those still using carbon-based neurons or otherwise—will believe them.

Some specific studies of the relationship between AI and religion include William Sims Bainbridge’s God From the Machine (2006), which investigates the question of whether religious activity might occur spontaneously in machine learning. Bainbridge and Stark (1987) proposed a general theoretical framework for the scientific study of religion, which included, among many other propositions, reasons for the emergence of cults and for cult affiliation; individuals characterized by high levels of education and social isolation, it is suggested, are more likely to participate in cults, a theory supported in part by practical testing (Bader and Damaris 1996). In 1995 Bainbridge employed the technique of neural network modelling to test his theory, and to identify the reasons for human religious behaviour. The idea of religion, however, is not part of the data supplied (484): “While not denying the possibility of God’s existence, our theory attempts to explain human religious behavior without assuming the truth of religion. Therefore, there is no axiom asserting the existence of the supernatural.” The program attempts to formalise communication and models a community who seek exchanges with each other in search of the basics of existence and of beneficial exchanges leading to personal rewards (486 emphasis original):

> In the scenario accompanying the program, they are called energy, water, food, oxygen, and life. Some people are producers of one or another of the first four rewards, and the simulation models the development of a little economy based on exchange of these consumable rewards. But none of the 24 people can provide each other with eternal life.

Bainbridge describes how, in the attempt to find suitable exchange partners for these more intangible rewards, the concept of the supernatural, within the context of a folk religion, may emerge spontaneously (492):

> In the subculture-evolution model, an intensely interacting group of individuals commits itself to the attainment of rewards, some of which are very difficult or even impossible to obtain. As they exchange rewards among themselves, they also exchange explanations about how to get other rewards, and in
the attempt to satisfy each other, they magnify slightly their positive evaluation of explanations. Those explanations that can be evaluated empirically will be rejected, leaving the nonempirical (supernatural) explanations that cannot readily be evaluated. Faith will spiral upward, and the group will create a folk religion through a series of thousands of tiny communication steps.

In a more developed version of the methodology, Bainbridge (2006) concludes that this is a natural, and to some extent inevitable, process that he is able to model in various ways. He posits that it is quite feasible that, primed with appropriate theological information, machines might also “communicate ideas as if they were exchange partners engaged in theological discussion with each other” (137), and that “once separated to some degree from external control, the evolving cult develops … the end point of successful cult evolution is a novel religious culture (139).”

7.0 Conclusion

The phenomenon of bias is found to be widespread in machine intelligence, both in data per se and in semantic content derived from text corpora. Most of the studies of machine bias have focussed on demographics such as gender, race, and occasionally, social class. Although it is mentioned in a few papers as another potential focus for bias, religious affiliation has not been investigated in the same way, despite the obvious existence of religious prejudice in society. However, examination of the well-established and influential resource WordNet shows high levels of bias in its structural associations and use of language, almost certainly as a result of inheritance from vocabularies used in its original construction. Because of its widespread use in the creation of search and discovery tools, particularly automatic classifiers, WordNet is likely to have passed on these prejudices.

Addressing the problems of bias has mainly concentrated on technical solutions, but recent research in modelling the ethical behaviour of humans suggests that intelligent agents may be able to develop a sense of fairness and moral responsibility independently of humans, and one not assuming pre-programming with specific rules. Several writers speculate that, in time, a sense of the religious could also emerge, and can provide a detailed demonstration of how this might happen using similar neural network methodologies. In time, robots themselves might be equipped to deal with the phenomenon of religious prejudice.

References


Fox, Melodie J. 2016. “Legal Discourse’s Epistemic Interplay with Sex and Gender Classification in the Dewey Decimal Classification System.” Library Trends 64, no. 4: 687-713.


Semantic Enrichment of Linked Personal Authority Data: A Case Study of Elites in Late Imperial China†

Shu-Jiun Chen

Institute of History and Philology, Academia Sinica, Taipei 11529, Taiwan,
<sophy@sinica.edu.tw>


Abstract: The study uses the Database of Names and Biographies (DNB) as an example to explore how in the transformation of original data into linked data, semantic enrichment can enhance engagement in digital humanities. In the preliminary results, we have defined instance-based and schema-based categories of semantic enrichment. In the instance-based category, in which enrichment occurs by enhancing the content of entities, we further determined three types, including: 1) enriching the entities by linking to diverse external resources in order to provide additional data of multiple perspectives; 2) enriching the entities with missing data, which is needed to satisfy the semantic queries; and, 3) providing the entities with access to an extended knowledge base. In the schema-based categories that enrichment occurs by enhancing the relations between the properties, we have identified two types, including: 1) enriching the properties by defining the hierarchical relations between properties; and, 2) specifying properties' domain and range for data reasoning. In addition, the study implements the LOD dataset in a digital humanities platform to demonstrate how instances and entities can be applied in the full texts where the relationship between entities are highlighted in order to bring scholars more semantic details of the texts.

Received: 30 September 2019; Revised: 14 November 2019; Accepted: 15 November 2019

Keywords: open data, linked open data (LOD), property, Database of Names and Biographies (DNB), information, semantic

† Presented at ISKO-UK 2019: The Human Position in an Artificial World: Creativity, Ethics and AI in Knowledge Organization, at City University, London, UK, July 15-16, 2019. The author would like to thank Dr. Cheng-yun Liu for his helpful advice on social history of the Ming and Qing Dynasties in this Paper, and Mr. Lu-Yen Lu, Ms. En-Ting Yu and Dr. Hsiang-An Wang for their assistance throughout the project. The work was supported by the Academia Sinica Center for Digital Cultures (Grant No. ASCDC-107-13), and the Ministry of Science and Technology (Grant No. MOST 106-2410-H-001-025).

1.0 Introduction

Semantic relations between entities are the basic units of knowledge organization (Green 2001; Stock 2010). This paper discusses the issue of semantic enrichment in linked open data (LOD) research. The study will use the Database of Names and Biographies (DNB) as an example to explore how, in the transformation of original data into linked data, semantic enrichment can facilitate research inquiries and enhance engagement in digital humanities. Hosted by the Institute of History and Philology, Academia Sinica (Taiwan), the DNB contains 35,666 records of Chinese historical persons who are cultural and socioeconomic elites in late imperial China (1368-1911), extracted from various historical archives for the purpose of supporting historians’ research. Each metadata record has information including name, alternative name, dates of birth/death, native place, biographical data, work experiences, related persons, specialties, academic background, job titles and references. Semantic enrichment is one of the current LOD project’s core research streams, developed by collaborating with historians for converting the DNB from legacy databases to linked data format for the purpose of digital humanities. The study has deployed the methods of data modeling, data reconciliation and data enrichment during transforming the legacy metadata records into LOD in order to add value that is structured in a machine-processable format and gives more meaning to the dataset (Hyvönen 2018; Van Hooland and Verborgh 2014; Zeng 2019). The DNB data model is composed of five
core classes (i.e., agent, event, place, object and time classes), sixty-seven properties from sixteen semantic vocabularies (i.e., bio, dbpedia, dcterms, gvp, leo, owl, rdfs, schema.org, skos, etc.) and reuses three external resources (i.e., AAT, VIAF, TGAZ). The LOD dataset contains more than two million triples and is available to query linked data with SPARQL from the LODLab of Academia Sinica (http://data.ascdc.tw/en/sparql.php). The following presents the preliminary results of the study, which we have defined instance-based and schema-based categories of semantic enrichment.

2.0 Instance-based semantic enrichment

Instance-based, or entity, enrichment, in the context of LOD is a process to endow the original data with other supplementary or supported knowledge, which is beyond the original data content, from external resources. This is done to extend the perspective of the data itself and also integrate heterogeneous resources into a more complete and sophisticated knowledge. By enriching the entities in the LOD-based dataset, it not only enlarges the knowledge base itself but also inspires new perspectives for further research and interpretation of the study results. There are three approaches identified in the study to achieve entity enrichment as follows.

2.1 Enriching the entities by linking to diverse, cross-domain external resources in order to provide additional data of multiple perspectives

In the DNB dataset, which is focused on the information of Chinese historical figures, the agent entity is linked to the cross-domain external resources. For instance, in the agent entity of Tseng Guo-fan (曾國藩, 1811-1872), a famous literati and high minister of the late Qing period, the AAT (Art & Architecture Thesaurus) concept of “calligraphy” is directly linked with the agent by the specialty property (dbpedia-owl:specialty). In the original DNB metadata, a specialty of Tseng is presented in literal as “書法” (calligraphy) in Chinese characters. Since the AAT is a multilingual thesaurus including English, Chinese, German, Dutch and more (Harpring 2018), the link to the AAT term for “calligraphy” (AAT 300053162) can enrich the agent entity in the related property with information on the same concept and its definition in other languages. This could enhance understandability for users who do not know the Chinese language.

2.2 Enriching the entities with missing data which is needed to satisfy the semantic queries

As to the information on the native place of a person, since its value shows the historical administrative name of a place and is not entirely equivalent with the current place name, we, therefore, added a new contextual entity for a place for each agent when the referred agent has information in the field of “native place.” The method of adding a place entity is not only to describe the historical name but also to be enriched by linking to the linked data of the Chinese historical place name in the Temporal Gazetteer (TGAZ) developed by Harvard University, which defines the geographical range and temporal information of the related historical place name. Such a place entity is linked to the agent entity of this study by the native-place property (ascdc:nativePlace) and also carries the label (rdfs:label) and the link to external resources by the same-as property (owl:sameAs) (shown in figure 1).

Another reason to reuse the terms of Chinese historical place names from TGAZ is that on the SPARQL (SPARQL Protocol and RDF Query Language) interface for DNB dataset, the study applied Chinese historical Geographic Information System (GIS) maps for showing an agent’s native

Figure 1. Enriching the entities with missing data, which is needed to satisfy the semantic queries.
place in certain examples of SPARQL query. For instance, when one queries “which provenances did the Qing agents ranked as jinshi (進士, “presented scholar,” the highest degree of the Chinese Imperial civil service examinations) come from?” Since the current Google Map as open source shows only the global map in modern political and administrative boundaries, it is quite different from the historical map. The location of a city from hundreds of years ago might also be different from its current geographic situation. Since each term in the Temporal Gazetteer contains the longitude and latitude information of a historical place name in different dynasties or periods, the application of TGAZ is reasonable for displaying the related geographic information in the historical GIS-maps.

2.3 Providing the entities with extension to a knowledge base

The extension of data content in the entities to a knowledge base is regarded in this study as one part of data enrichment. It could not only enable the entities with more detailed information from the external resources but also make linkage of the original entity to an entire knowledge base of a certain subject field, which can inspire new perspectives for further research and interpretation of the study results. For instance, the study makes extension of entities to the time ontology, which is one part of the knowledge base. In the current LOD-based dataset of the DNB, the entities relating to the temporal information on the official career of a person is particularly extended to the “term lists of the time and periods,” a controlled vocabulary developed by the study for defining the Chinese historical periods and yearly times of all dynasties in China (see Figure 2). The application of such term lists is based on the “time ontology,” which is developed by the study and based on W3C’s time ontology in OWL (Cox and Little 2017).

For example, the DNB person agent Ding Bao-zhen (丁寶楨) was appointed as Governor of the Sichuan Province (四川總督) between 1881 and 1886. In the data model design of DNB, a person agent’s official career in the government is expressed as entities of official service (asdc:OfficialService). To describe the beginning date and end date of Ding’s appointment, this entity for official service is linked to an entity of temporal period (time:TemporalEntity) by at-time property (leo:atTime), which further describes the beginning and end date of the related period for official service in the historical Chinese era year, represented by the “time:instant” entity. To extend the knowledge base for that temporal information, the study’s time ontology is applied to describe the hierarchical temporal details of the instances as the “7th year of Guangxu” (光緒七年, 1881) and “12th year of Guangxu” (光緒12年, 1886) and linked with the entity for era name (光緒), emperor name (光緒皇帝) and dynasty name (清朝) in “time:propertInterval” by the properties “time:interval-During” and “time:inside” (see Figure 3).

After extending the entity with the temporal terms to describe the beginning and ending year of a certain period of an official position, the information of these mentioned historical years is expressed as entities of thing and can further be linked to the era names, emperor names and
the dynasty names in Chinese history. Therefore, a hierar-
chically structured knowledge base, which is focused on
Chinese historical temporal names, is entirely integrated
into the dataset and enriches the data content of each re-
lated person agent.

3.0 Schema-based semantic enrichment

Schema- or property-based enrichment in LOD is a pro-
cess to enable hierarchical or associative meaning within
pairs of property, which could create a relationship be-
tween related entities and also enable a meaningful and ef-
ficient data query in a hierarchical semantic structure.

3.1 Enriching the properties by defining the
hierarchical and associative relations between
properties

To a certain extent, the applicability of enriching the prop-
erties in the LOD datasets depends on whether a hierar-
chical or associative meaning exists between different data
elements of the original metadata. In the data element of
the current DNB datasets, such related meaning is especially
found in the data element as the “personal relations”
(人物關係) of an agent, in which a person’s connections to
another agent is linked and expressed by reusing suitable
properties. As an example, the semantic relationships of an
agent A to his teacher, agent B, is defined as the has-teacher
property (agrelon:hasTeacher), while the relationships of
agent A to his grandparent, agent C, is expressed as the has-
grandparent property (agrelon:hasGrandparent).

In fact, the personal relation between agents is a mutual-
or hierarchical-expressible relationship. If A is stu-
dent of B, then the B should be the teacher of A. In the
semantic data model design, such mutual or hierarchical
relation can be defined by enriching the definition of the re-
lated properties. In the current world of semantic web,
RDFS and OWL are the two types of data vocabularies,
which are mostly applied to enrich the relations between
properties as the abovementioned cases and also to enhance
In particular, the subproperty-of property (rdfs:subpoper-
ytyOf) can be used to mark the hierarchical relation between
properties, while the inverse-of property (owl:inverseOf) is
suitable to describe the mutually affected relations between
properties on the same level (see Table 1).

Taking the agent Zeng Guo-fan (曾國藩, DNB: NO000000058) in the DNB datasets as an example, the fig-
ure is linked to the agent Li Hong-zhang (李鴻章, DNB: NO000002242) by the has-student property (“agrelon:has-
Student”). Since the relationship between a teacher and stu-
dent is mutually referred, if the inverse-of property (owlin-
verseOf) is reused and enriches the definition of the has-
student property (agrelon:hasStudent), the reverse relation
expressed as has-teacher property (agrelon:hasTeacher) would also be findable by data reasoning.

![Figure 3](image.png)

**Figure 3.** Extension of the entity with knowledge-based external resource (Example: Extension of the entity for official service with hierar-
chical information on the beginning and end date of Ding Bao-zhen’s appointment as Governor in Sichuan between 1881 and 1886).

<table>
<thead>
<tr>
<th>Properties</th>
<th>Domain</th>
<th>Range</th>
<th>Function of property enriching</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>rdfs:subPropertyOf</td>
<td>Property</td>
<td>Property</td>
</tr>
<tr>
<td>2</td>
<td>owl:inverseOf</td>
<td>Property</td>
<td>Property</td>
</tr>
</tbody>
</table>

*Table 1. Enrich the properties by using rdfs:subpropetyOf and owl:inverseOf.*
With the same example of Zeng Guo-fan, the figure is further linked to the agent Yuan Bingzhen (袁秉楨, DNB: NO000012514) by the has-child-in-law property (agrelon:hasChildInLaw) and to Zeng Guangquan (曾廣銓, DNB: NO000008193) by the has-grandchild property (agrelon:hasGrandchild). Since those different types of relations to a person can be clustered in a broader range of properties such as relatives, a hierarchical structure of property can be hence defined by using the sub-property-of property (rdfs:subPropertyOf) to structure the has-child-in-law property (agrelon:hasChildInLaw) and the has-grandchild property (agrelon:hasGrandchild) both under the property as has-relative (agrelon:hasRelative).

3.2 Specifying properties’ domain and range for data reasoning

In the semantic web, each data can be expressed as a triple, which is composed of subject, property and object. From them, the major function of a property is to enable the entities of information (subject and object) with a semantic relation, which could enable the data query in a logical, machine-processable and machine-understandable way. In other words, the property plays a role as the bridge to connect the subject with object and thus construct complete, meaningful information in the data. However, the use of a certain property is not arbitrary. Each property has its own definition to which condition or restriction can be applied to link the subject- and object-entities.

In the current data model design for the “Database of Names and Biographies” (DNB), sixty-seven properties from sixteen vocabularies are reused to describe the biographic information on Chinese historical figures and relations between figures. The information on the domain and range of all reused properties is described in the specification of DNB ontology, seen in the selected examples in Table 2. Such specification can be used as referential standard by the semantic data model design and also defines which data context that a property could be reused in the data structure.

In the semantic data model design, the domain of a property is always the instance of an entity (or a class), which defines the subject of an information. As in the aforementioned example, a property as “bio:father” in the DNB is applied to describe the information on the father of a person. The domain of this property is defined as “foaf:Agent,” which means this property is only suitable to be applied by an entity of the agent. In the machine-processable form, it could be formulated as “rdfs:domain foaf:Agent.” However, the range of a property could be expressed as a different data type, such as the instance of an entity (or a class), literal information, date decimal numbers or quantity of item. Again, as in the example (Table 2), the specification also defines the range of “bio:father” as “foaf:Agent,” which means that object information should also be an instance of agent entity. In the machine-processable form, it could be formulated as “rdfs:range foaf:Agent.”

4.0 Applications of LOD for digital humanities

The input of the LOD-based dataset of the Database of the Names and Biographies (DNB) into the system of the Digital Humanities Research Platform (DHRP), developed by the Academia Sinica Center for Digital Cultures (ASCDC), is a linked data application to enhance the reusability of the DNB data. Additionally, the study demonstrates the possibility of applying a LOD-based dataset to enlarge the research scope of the scholars in digital humanities and to integrate into digital research tools—using examples in the DHRP.

The DHRP is an open, cloud-based text repository to enhance the research of digital humanities, which is developed as a platform for online services based on the needs of scholars. The platform is equipped with different digital tools for text and visual analytics, such as text annotation,

<table>
<thead>
<tr>
<th>Property</th>
<th>URI</th>
<th>Label</th>
<th>Type</th>
<th>Comment</th>
<th>Domain</th>
<th>Range</th>
<th>Quantification</th>
<th>Data type</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>bio:father</td>
<td><a href="http://purl.org/vocab/bio/0.1/father">http://purl.org/vocab/bio/0.1/father</a></td>
<td>Father</td>
<td>Property</td>
<td>To describe information on father of a DNB Agent Entity</td>
<td>foaf:Agent</td>
<td>foaf:Agent</td>
<td>0-1</td>
<td>Concept/ASDC</td>
<td>Liu yun father Liu Ton-hsung</td>
</tr>
</tbody>
</table>

Table 2. Specifying domain and range of the father property (bio:father) in DNB.
text similarity comparison, N-gram analysis, historical spatiotemporal visualization or social network analysis. The digital content of the DHRP is currently uploaded with texts from rare Chinese books for a total of more than 220 million words. In the current stage, the DNB dataset is already in the test version of the DHRP-platform. In particular, we use the DNB’s data on properties of the person’s relations, specialty and native place as practical cases of studies to map the text passages in the Qing Shilu (the Veritable Records of the Qing Dynasty/清實錄) and to demonstrate the semantic relations between different person agents or agent entities with place or concept entities in the historiographical works of the DHRP-platform. In total, more than 20,000 named-entities from the 93,431 text passages in the Qing Shilu are matched with the instances of entities in the DNB.

In the DHRP-platform, the mapped DNB personal names in the Qing Shilu will be marked up in different colors according to their types in the data unit of a triple in the DNB dataset. For instance, the person’s name belonging to the subject in a triple will be highlighted in blue, while names of an object in a triple will be shown upon a gray background (Figure 4). The type of semantic relationships (properties) between different agents will be represented in a dotted line, which link the subject entity with its related object entity.

When moving the mouse cursor onto the subject agent in the text, the platform will automatically present the related agents of person names in green with the type of semantic relation. Further, clicking on those names will direct to the website of the LOD-datasets, showing the data content of the related records of the persons (Figure 5).

For further presentation of the related data in DHRP-platform by using the tools for data visualization, the function of social network analysis (SNA) is integrated into the system to show the matched persons in Qing Shilu based on the DNB dataset (Figure 6).

In the original DHRP-platform, the scholars could only execute the text retrieval and comparison based on the literal context uploaded in the system. Users could not find out detailed information or definitions of the retrieved words or text, since the context was not linked to the external resources by a semantic method. After uploading the LOD-based dataset in the originally text-based DHRP-
Figure 5. Linking the matched person's name in *Qing Shilu* to the external resource in the DNB (Example: Moving the cursor onto the person name of Zeng Guo-fan (曾國藩) and showing his related persons retrieved in *Qing Shilu*; clicking one of the names as Zeng Guo-chuan (曾國荃) and linking to the equivalent resource in the DNB).

Figure 6. SNA-analysis showing the relation types between matched persons in a passage of *Qing Shilu* in different forms of data visualization. 

(1) SNA by e-chart; (2) SNA by D3.js data visualization.
platform, a detailed definition of the matched text (for example, the further biographical information of a person) can be shown in DHRP by reusing the related data in DNB dataset. This is accomplished through the named-entity recognition, which is a mapping procedure of the terms in DNB with the text in the platform. The semantic relations of a person to another person, place or concept will also be notified by endowing the type of relations (properties) in the platform. These could extend not only the knowledge base of a scholar but might also offer other relevant information or inspire research angles, which one might not take notice when retrieving the results merely in a literal context.

References:


Knowledge Organization, Data and Algorithms:
The New Era of Visual Representations*

Viviane Clavier

GRESEC—Université Stendhal Grenoble 3—Institut de la Communication et des Médias, 11, avenue du 8 mai 1945—BP 337—38434 Échirolles Cedex,
<viviane.clavier@univ-grenoble-alpes.fr>

Viviane Clavier is Associated Professor and Research Director in information and communication sciences. She is a member of Gresec at the University of Grenoble Alpes and her research focuses on knowledge organization and information seeking in the field of specialized and professional information. These studies focus on the themes of health and nutrition.


Abstract: This article shows how visual representations have progressively taken the lead over classical language-based models of knowledge organization (KO). The paper adopts a theoretical and historical perspective and focuses on the consequences of the changes in the volume of data generated by data production on the KO models. Until now, data visualization tools have been used mainly by researchers with expertise in textual data processing or in computational linguistics. But now, these tools are accessible to a greater number of users. Thus, there are new issues at stake for KO, other professions and institutions for gathering data that contribute to defining new standards and KO representations.

Received: 30 September 2019; Revised: 4 November 2019; Accepted: 4 November 2019

Keywords: knowledge, data, visual representations, images, information


1.0 Introduction

This contribution is concerned with the place of languages in the process of collecting, representing and organizing knowledge in a context characterised by the massive data production. By suggesting in the title that visual representations are becoming crucial in representing knowledge, we implicitly mean that they have supplanted languages. We are not saying that languages have disappeared, as the use of controlled languages is still relevant to heritage institutions and several fields of specialized information. However, the attention paid to languages has shifted to other knowledge objects (algorithms), other semiotics (data visualization), other levels of representation (textual data), other purposes of knowledge organization (browsing rather than categorization) and other properties (transparency and simplicity rather than precision and exhaustiveness). It is important to notice that the term “language” is not mentioned in the last report published by the IFLA (2018), which is revealing about new trends in information concerning libraries. This paper, therefore, raises a series of questions, puts forward certain hypotheses, looks into the reasons why languages are being called into question, and—if they are being replaced—by what and how. This study has a diachronic rather than historical dimension, which can be seen via research work in information and communication sciences (CPDirSIC 2018), and it also makes reference to our own research in which language, texts and discourse have always played a central role (Clavier 2014). Our presentation is in two parts: the first part shows how visual representations have progressively taken over from language-based representations in knowledge organization systems. The second part shows how this trend observed in research has also spread to civil society and to professionals concerned by information and communication. However, it is clear that languages are still very present, even if they appear to be relegated to a background descriptive function.
2.0 The challenges of knowledge organization: development and permanence

2.1 Applied perspective: languages for knowledge organization systems

The notion of “languages” will be restricted to symbolic systems chosen to encode units of language and meanings, which are used to feed knowledge organization systems (KOS), “tools intended to organize recorded knowledge resources” (Sosinisko-Kalata 2012). KOSs include “all kinds of organization schemes ranging from simple alphabetical or slightly structured lists (authority records, glossaries, dictionaries, parts lists, etc.) to hierarchical classification schemes (classification plans, general or specialised classifications, taxonomies, subject headings lists, etc.) or organizations emphasising relations that are not exclusively hierarchical (thesauruses, semantic networks, ontologies, etc.)” (Polity et al. 2005, 13). This perspective places us immediately in what should be referred to as “the narrow sense” of knowledge organization, namely the creation of information products intended to classify, organize and structure knowledge in order to facilitate access to information. This perspective is situated between a “broad” and theoretical conception of knowledge—for an ontological (Gnoli 2008) or social (Hjorland 2015) conception of knowledge—and an “intermediate” approach centered on the applications and forms of knowledge mediation that can be observed in professional situations (Clavier and Paganelli 2013, 2015).

What is common to documentary languages, classifications, folksonomies, ontologies, etc. and today’s “data” is that they use processing methods in order to represent content. Thus, at any point in time, knowledge is the result of choices and transformations governing its selection, perhaps its standardization and its categorization. The huge quantity of data to be processed is nothing new—even in the 1970s the development of document processing was based on the idea that automation would help reduce the costs of manually indexing the constantly increasing documentation that professionals could no longer handle (Chauvet 1982). What we intend to question here is the nature of the knowledge selected with regard to computer processing choices.

2.2 From the 1980s to today: the end of languages?

From 1980 to 1990, the development of full text information systems and other applications such as linguistic technology-based knowledge extraction (Chaudiron 2007), enabled the analysis of languages from the standpoint of linguistics (Van Slype 1982). The encounter between computational linguistics and documentation (Rouault 1987) brought together “these two disciplines even though they were separated by a wall” (Van Slype 1982, 87). Viewed from the logico-syntactic perspective of knowledge that could be processed by computer programs, natural languages were the subject of theoretical debates and considerable socioeconomic stakes. The questions raised were as follows (Lallich-Boidin et Maret 2005): 1) What is the relevant unit of meaning: the word or the morpheme? 2) What advantages and disadvantages do free languages have in comparison with controlled languages? and, 3) How can meaning be represented accurately while taking into account ambiguities and implications? It is necessary to use external resources in order to define indexing languages (thesauri, controlled vocabularies) or to the documents themselves—a question that raised the issue of faithful indexing in light of cost savings in order to maintain lexical resources. This approach conceived languages as being worked with and by computational linguists, and questioned in relation to language units. It focused on a detailed description of language, based on a compositional approach to meaning led by syntax in the form of logical representations, and was replaced by another approach based on an empirical representation of meaning in which the aim was no longer to understand the language but to describe textual corpora.

The 1990s thus represented a turning point characterised by the web development, paving the way to corpus linguistics (Habert et al. 1997), an approach that shifts the focus towards the creation and description of textual data. And so began the period of “large corpora” intended for annotation by exploration tools as a means of assessing the representativeness of language phenomena. From this standpoint, emphasis was again put on the “vast” character of data and text annotation or “tagging” aimed to collect typologies based on correlations of linguistic features (Biber 1993). These annotation methods, which were often manual or semi-automatic, combined with more robust information processing methods such as learning-based text mining, aimed to perform supervised or unsupervised automatic classification tasks. The knowledge resulting from corpus annotation can be morpho-syntactic categories (Poudat et al. 2006) or collocations (Tutin 1997), with phraseology having benefited from the joint development of symbolic methods and probabilistic methods via automatic learning. Thus, unlike knowledge produced by computational linguistics, which aims to automate applications, the knowledge resulting from annotations produced by corpus linguistics “is based on an iteration between the analysis of computer outputs and human consultation of texts or fragments of texts, for example concordances” (Valette and Egle 2014). The production of knowledge thus forms part of a tool-based approach aimed at exploring data in order to pinpoint linguistic recurrences (idioms, expressions, etc.). This perspective led to the development
of methods based on “opportunistic” linguistic approaches, i.e., those that exploit automatically identifiable surface phenomena.

Over time, the empirical approach became entrenched and enthusiasm for numerical processing methods increased. The development of data analysis tools and the generalisation of methods intended for social sciences and the “general public” was accompanied, as Dominique Boullier points out (Boullier 2015), by “the popular belief that announced the end of science and scientific theories.” Thus began a new period in which big data “would be the effective measurements of reality” (ibid.). According to this view, data are based on the collection of bags of words obtained from the web that are characterised by their volume and can be visually represented in two-dimensional spaces. It is then a matter of interpreting and discovering this knowledge, the status of which is not always very clear; does knowledge represent themes, subjects, key words, discourses, lexicometric universes, etc.? The drawback of these tools often lies in the opacity surrounding the data pre-processing stages, the choice of categories (both lexical and textual) or units of discourse. As pointed out by Pascale Sébillot in 2002, mixed methods have indeed been developed, but in limited fields where the aim is no longer to understand texts in detail but to obtain representations of meanings that are useful for precise applications (Sébillot 2002). Thus, rule-based systems are used to develop resources for analysing sentiments and opinions (Poibeau 2014), so that it is possible to annotate data intended to train classifiers. Other applications also rely on the profusion of data, such as machine translation, data journalism and fact checking. However, as pointed out by Thierry Poibeau (ibid.), the part connected with computational linguistics is variable, and “the recurrent difficulties concern system adaptability, the time required to develop resources for a new field and the availability or lack of a sufficient number of examples.” The researcher also indicated that other frequent questions concern the definition of the information sought and the quality of the results obtained.

For a number of years now there has been considerable interest in visual representation, as revealed by a recent review of works on the subject carried out by researchers at the University of Swansea (Rees and Laramée 2019). The authors’ research goes back as far as 1967, to the Semiology of Graphics by Jacques Bertin, one of the first monographs devoted to the question. According to them, the following forty years were not particularly prolific, but since the early 2000s there has been a plethora of publications, to the point where it is difficult to keep track of them (610). Visual representation must be understood as both a process of creating visibility and as the result of that process. In the former case, it is “the result of the instrumented exploration of masses of data that become suitable for generating indicators, maps, etc.” (Reymond 2016, 11). In the latter, the produced visual representations form the material for computer graphics, laid out in such a way as to form a language in its own right. However, what is visualized may be data, information or knowledge. “Data visualization” is the generic term used to designate the tools of “dataviz,” involving “a semiotic transformation between the results of a data analysis (numerical, categorial, textual) and a graphic representation” (Hachour 2015). The “visual representation of information” is a branch of computing that “uses visual and interactive representations of abstract data on a computer to amplify cognition” (Lamy 2017, 76 ff) and has six possible goals: displaying a large quantity of data, facilitating the search for given information, detecting patterns, enabling visual inferences, monitoring the occurrence of events and exploring data sets. The graphic representations obtained in this way serve several applications intended to visualize information, such as journalistic or documentary information. Thus, data visualization for information (Yikun et Zhao 2016) is a means of discovering journalistic news by establishing “new interactions” and presenting them visually. As far as documentary information is concerned, it may be visualized by maps that help in browsing classifications (Dewey), directories (Rameau) or catalogues (Papy 2005). Lastly the “visual representation of knowledge” falls within the field of research into artificial intelligence and is based on the representation, modeling and visualisation. According to the computer science researcher Jean-Baptiste Lamy (Lamy 2017, 12), a distinction must be made between the iconic visual representation of knowledge, which involves translating knowledge by using an iconic language including a pictogram glossary and a grammar, and a structural visual representation of knowledge, which involves representing the structure of knowledge graphically using the techniques of visual representation.

3.0 Knowledge dissemination: priority of visual representations for many actors

The choice of verbal or visual semiotics to represent knowledge in information systems is not simply a matter for science but falls within a much wider spatio-temporal, socio-economic and cultural context. Visual representations are now the predominant modes of expression in our society, as evidenced by computer graphics, fixed and animated images (videos) in the field of public and scientific communication, in the media and on the web. To remove any ambiguity, it is not a question of demonstrating that we live in “a world of images” as suggested by the philosopher Franck Robert in his introduction to the joint publication “Philosophies de l’image,” but rather that the importance assumed by visual representations—or “visuals”
as they are referred to by communication specialists—as forms of semiotization influences the ways in which knowledge is represented and organized. The rest of this article will present a few trends observed among actors who, for one reason or another, play an important role in the dissemination of knowledge representation standards or formats. These examples—which are not derived from a reasoned corpus—illustrate the fact that knowledge is always rooted in particular periods and places and probably depends on fashions.

3.1 Promotion of images in web publication tools and techniques

Images are promoted by the web professionals who define the writing standards and techniques aimed at optimising the visibility of websites via their referencing and positioning in search engines. There are very many sites aimed at web publishers giving recommendations on how to write for the web and setting out rules on concision, simplicity, content structure, addresses, etc. With regard to images, they may be fixed or animated, such as videos. On the website Annei, images and text, “always considered to be rivals,” are now presented as complementary, subject to a few adjustments intended to prevent images being restricted to a secondary role. The site advises combining “alternative text,” a “concise, descriptive textual equivalent” along with images that do not exceed “250 characters,” or “ensuring that video soundtracks will be made accessible by a retranscription in text format that can be detected by search engines.” As far as the choice of images is concerned, it also gives recommendations on their content: “Choose images that have an informative value.” This advice is aimed at eliminating images that are purely illustrative. Images considered to be informative include: “computer graphics, diagrams, photographs taken in real situations, providing more than a visual taken from a database and is of purely decorative value” (ibid). In this last instance, it advises adding “a concise legend that gives meaning to the image, whenever possible,” in order to reinforce “the image’s impact rating.”

The advice given with regard to publishing on the social media is that images should even replace text on the grounds that an image simply requires a click to score a hit, thus guaranteeing traffic that generates profits or viewership. This recommendation appears to be connected with users’ observed habits: they click on images to browse the internet:

Pinterest and Facebook users are accustomed to zigzag browsing. One click leads to another click, an image leads to a link where there are other images, which also lead to other links. Images are like Tom Thumb’s little pebbles. If they are scattered intelligently, they will attract far more traffic to a website than a simple ‘quote.’

When you realise that photos posted on Facebook can generate 53% more ‘likes’ on a post, you make the most of them. Users are far more inclined to like or follow a brand if it is active and shares images. Textual promotion is a thing of the past.

Web writing standards for the social media now advise choosing visual semiotics: it is better to “show” than to “tell,” words are “there to underscore the image.” “Visual strategies are to be used to win visitors,” etc. Even so, the eternal question of document description, representation and indexing is still raised in the context of information searches in the form of user requests. Ultimately, then, images always lead back to the question of languages and the choice of language to describe them. This question is dealt with more specifically on sites intended to enhance the visibility of websites in search engines (search engine optimization) or on the social media (social media optimization) or specifically in engines dedicated to marketing (search engine marketing). Reference is made to W3C, which defines web languages and standards and gives the image formats accepted by search engines, description tags and advice on describing the alt and title attributes for presenting “the content of an image clearly and concisely.”

The alternative text of images: Google also references images! Always remember to complete the alternative text (alt tag) of the images in your articles. Introduce your main keyword within a short expression. The alt tag regularly omitted is that of the image on the first page, which weakens image referencing. To avoid this, fill in the Alternative text field when you define your first page image.

It is clear that images, far from replacing text, once again raise the question of textual representation, the choice of words and their descriptive virtue, reversing the roles of “illustrative” images in favour of text. From the point of view of information searching, these trends have led to the development of research into the automatic classification of images, facial recognition, and the production of languages for describing and annotating images, such as ontologies. Indeed the “Google Image” search engine is becoming increasingly popular with users to the extent that natural image referencing now appears to be a priority.
3.2 Development of data visualization tools for the general public

Hitherto reserved for specialists in multidimensional descriptive statistics, certain methods such as principal component analysis (PCA) and correspondence factor analysis (CFA) have been widely used in the humanities and social sciences to study literary or political texts from different viewpoints: lexicon, vocabulary, style, themes, etc. An extensive literature devoted to textual statistics has been built up since the late 1950s, the common theme of which is the production of knowledge organization systems in the form of verbal semiotics: dictionaries (Guiraud 1959), indexes, concordances, repeated segments, textual time series, parallel corpora, etc. (Salem et Lebart 1996). In these works, visual representations—entire lexical tables, figures representing factorial planes, etc.—are not a goal in themselves but are considered to be an “aid in reading a series of texts” (135 ff). These introductory works give guidelines on: “How to interpret distances,” “Examples of applications,” “Reading a figure,” etc. It is thus clear that the interpretation of textual data is subject first of all to lexicometric reorganization in the form of visual representations, which are the result of a set of processing operations and choices concerning the size of the textual corpora, their representativeness and segmentation, the definition of counting units, etc. As a follow-up to these methods, text mining includes statistical data processing and tasks such as supervised or unsupervised classification (Ibekwe-SanJuan 2007). Developed by computer scientists specializing in learning, these methods approach texts as data intended to train classifiers, whose scores are measured and compared. Applied research as information retrieval, pattern detection, scientific monitoring, etc. are always envisaged, so that the text in the strict sense and considered in its production conditions, is not the focus of attention. It is this transition from viewing text as sacred to viewing it as an instrument, a training corpus, that appears to be at the origin of an interpretation centered on understanding trends, regularities, mass effects, etc. and less on what is of the order of the invisible, which can only be retrieved through knowledge of the texts.

At the same time, dataviz tools are becoming far more commonplace as many actors require analyses of discourse on the social media, blogs, Twitter, forums, etc. Thus, for economic reasons, survey institutions have introduced so-called passive methods, inspired from big data, which are qualitative, can be automated and used to complement quantitative surveys. There is a wide social demand: marketing, political polls, consumption, etc. There is a multitude of websites that list dataviz tools, distinguishing between various families: mental maps, visual representation of relational data networks, information mapping, etc. Their success would appear to be due to their transparency; simplicity and faithfulness to the data, as can be seen in this example of an announcement presenting the advantages of dataviz for representing textual or numerical data.

Dataviz, or data visualization, is a practice that we encounter on a daily basis without realizing it, just by opening a newspaper or watching the television. The simplest example is the survey. With the digital age, it has become a powerful communication tool.

Dataviz: a definition that is easy to understand.

In a society that is increasingly attracted to graphics, data visualization takes precedence over raw data. It helps to throw light on information that is apparently complex or is submerged in a large quantity of parameters. The term dataviz thus refers to a set of visual representations of these raw data. What is its principal objective? To throw light on a phenomenon by giving it a more ergonomic appearance than a spreadsheet filled with figures.

Transparency refers to the immediately perceptible and understandable character of a message when it is represented in a visual form, which would give it undeniable communicational virtues. Simplicity is the ability of the tools to provide a summary representation in visual form of information that is expressed in a verbal manner, which is thus felt to be more complex. Lastly, faithfulness is the ability of the tools to represent data such as that emanating from individuals without any form of mediation. These properties, which are presented as technical assets, are the subject of ongoing theoretical debates on the differences in semiotic status of pictorial and written signs (Christin 2012). These arguments hark back to the sterile debates on the more transparent character of pictorial signs in comparison with written signs on the grounds that there is no need to use an arbitrary code to interpret an image, in contrast to a text; likewise simplicity no doubt refers to the small number of signs used, as a small set of visual forms is always more concise than an infinite number of verbal signs combined to describe a language. These debates are far from over, and they question the pertinence of knowledge encoding processes without taking into account reception conditions. Lastly, faithfulness is also a question that is widely discussed in the information-documentation field with regard to the merits of free and controlled indexing languages.

The representation of knowledge in visual form is also of interest to communication specialists who approach information design from the aesthetic standpoint. They are required to possess new dataviz skills at the crossroads between computing, statistics and communication (Arruabar-
Supported by specialist disciplines in big data processing and encouraged by research and education policies, data visualization is more than just a tool for the knowledge management profession; it is a genuine aid to discovering knowledge. It is no longer a matter of naming knowledge in order to understand it, but rather of showing it in order to make it visible, a perspective that leaves considerable leeway for new actors, new specializations and new data collection structures that are helping to redefine standards, forms of writing and methods for representing knowledge.

Notes
1. The examples of sites mentioned here are among the top results obtained by Googling “web image and text publication.”

References

Figure 1. Screen capture of the exhibition devoted to the visual representations of Paul Otlet, obtained from GoogleArts&Culture.


Ifla. 2018. Riding the waves or caught in the tide? Navigating the evolving information environment. Insights from the trend report.


Data Papers as a New Form of Knowledge Organization in the Field of Research Data†

Joachim Schöpfel*, Dominic Farace**, Hélène Prost***, and Antonella Zane****

*University of Lille, GERiiCO laboratory, ORCID 0000-0002-4000-807X, <joachim.schopfel@univ-lille.fr>
**GreyNet International, Amsterdam, ORCID 0000-0003-2561-3631, <dominic.farace@textrelease.com>
***CNRS, GERiiCO laboratory, ORCID 0000-0002-7982-2765, <helene.prost007@gmail.com>
****University of Padova, ORCID 0000-0001-7218-6068, <antonella.zane@unipd.it>


Abstract: Data papers have been defined as scholarly journal publications whose primary purpose is to describe research data. Our survey provides more insights about the environment of data papers, i.e., disciplines, publishers and business models, and about their structure, length, formats, metadata, and licensing. Data papers are a product of the emerging ecosystem of data-driven open science. They contribute to the FAIR principles for research data management. However, the boundaries with other categories of academic publishing are partly blurred. Data papers are (can be) generated automatically and are potentially machine-readable. Data papers are essentially information, i.e., description of data, but also partly contribute to the generation of knowledge and data on its own. Part of the new ecosystem of open and data-driven science, data papers and data journals are an interesting and relevant object for the assessment and understanding of the transition of the former system of academic publishing.

Received: Revised: Accepted: 3 December 2019

Keywords: data, papers, journals, research, metadata

† Presented at ISKO-France 2019: Open Data and Megadata in Social Sciences: New Challenges for the State and Organization of Knowledge?, at Université Paul-Valéry, Montpellier 3, France, October 9-11, 2019. The paper builds on the scientific and professional
experience of all authors in the framework of the international Grey Literature Network Service and its international workshops on data papers (Netherlands, Czech Republic, Italy, and the United States), its collection of data papers in the Dutch EASY data repository, and its follow-up study of enhanced publications.

1.0 Introduction

In the context of open science, an increasing volume of research data are made available on the internet, contributing to the so-called big data of science. New tools, methods, and infrastructures have been developed for the dissemination, processing, analysis, and preservation of research data. Data papers are part of them.

Data papers are a young species of academic publishing. In 2006, Pärtel stated (99) for the field of ecology that “until now … very few data papers have appeared.” In fact, most of the data papers or papers about data papers have been published since 2008 and 2009. Yet, as Smith (2012, 16) reminds, “the concept has actually been around for quite a while (even if) the older journals that date from the print era tend to be not particularly useful in the modern environment.” In fact, one (the first?) data journal (Journal of Chemical and Engineering Data From ACs) was already launched in 1956 (see the timeline in Garcia-Garcia et al. 2015).

The simplest definition (Callaghan et al. 2012, 112) is that data papers focus on “information on the what, where, why, how and who of the data” rather than original research results.

Data papers have been defined as “a searchable metadata document, describing a particular dataset or a group of datasets, published in the form of a peer-reviewed article in a scholarly journal.” They are published in specific data journals like Data in Brief (Elsevier) and Scientific Data (Nature) or in regular academic journals with special sections for data papers, like BMC Research Notes GigaScience (Oxford University Press) and PLoS One. Most data papers are published in journal platforms; yet, some are (also or exclusively) published on data repository platforms. Unlike usual research papers, the main purpose of data papers is to describe datasets, including the conditions and context of their acquisition and their potential utility, rather than to report and discuss results. Also, it is generally assumed that data papers are short papers with up to four pages.

In the “classical” research paradigm, the focus is on articles presenting results while research data are useful for the validation of published research findings. Data papers invert the roles, insofar the paper’s main function is to inform about and link to research data on data repositories, contributing to their findability and reusability. Are data papers complementary to research papers, or will they replace them, as a seamless and direct way of providing access to research results?

Also, traditional knowledge organization makes a clear distinction between research results (datasets), the analysis and discussion of these results (papers), and the description (cataloguing, abstracting, and indexing) of those datasets and papers. This emerging category of data papers appears to challenge this clear distinction, interlinking datasets, papers and metadata, blurring boundaries, changing priorities, and modifying the basic purpose of academic publishing.

Built on an overview of recently published studies, the following study produces an empirical update on the publishing of data papers: the number and development of data papers and journals, the country and language of publications, the platforms and publishers, as well as the business models. The purpose of our paper is to analyse data papers as a new tool of scientific communication and to produce insight on their contribution to the organization of scientific knowledge via questions pertaining to the production and the functions of data papers:

- How are they “written”? 
- Which is the link with data repositories, metadata, and other papers? 
- Which is the (potential and real) part of automatic or semi-automatic production? 
- Which is the part of human added value? 
- Which degree of standardization, which link between metadata formats and the data journals’ author guidelines? 
- In which way are data papers related to the so-called “FAIR Guiding Principles for Scientific Data Management and Stewardship”? (Wilkinson et al. 2016)? 
- Do they just improve the referencing of datasets on repositories, or do they fulfill other roles? 
- Are data papers “written by machines” and meant to be “read by machines”? 

The paper will conclude with a conceptual approach to data papers as part of the organization of knowledge based on research data in the context of open science.

2.0 Literature overview

2.1 Definitions and functions

An increasing number of journal editors announce the launch of a new section with data papers. They put forward different objectives, even if the main purpose is similar: to inform about research data and to foster their accessibility and reuse. Three examples among others illustrate the diversity of goals:
The objective of *The International Journal of Robotics Research* is “to facilitate and encourage the release of high-quality, peer-reviewed datasets to the … community” (Peter and Corke 2009, 587).

*Studies in Family Planning* tries to promote “interdisciplinary research and integrative analyses by making accessible to researchers, policymakers, students, and donors data that may be useful in answering critical questions of interest to … readers” (Friedmann et al. 2017, 291).

The French journal of information and communication sciences *RFSIC* invites data papers to describe the scientific process, methods, and tools that result in research data in a Bruno Latour perspective, “since they never just magically appear” (Le Deuff 2018, 2).

The publisher Pensoft describes (Penev et al. 2012) a data paper as “a scholarly journal publication whose primary purpose is to describe a dataset or a group of datasets, rather than to report a research investigation. As such, it contains facts about data, not hypotheses and arguments in support of the data, as found in a conventional research article.”

The term remains ambiguous. For instance, Bordelon et al. (2016, 1) define data papers as “papers that present, analyze, or use data obtained with the respective facilities” (i.e. observatories). Pärtel (2006) considers data papers as a kind of “abstracts” that aim to collect, organize, synthesize, and document data sets of value in a given field; only the abstract appears in a data journal (or the data paper section of a regular journal) while the data and metadata are available through a field-specific data repository on the internet. For Penev et al. (2012), their purposes are three-fold: “to provide a citable journal publication that brings scholarly credit to data publishers; to describe the data in a structured human-readable form; and (and) to bring the existence of the data to the attention of the scholarly community.” At first sight, data papers, in spite of their common general purpose, appear to belong to a rather heterogeneous and dissimilar new kind of document. Our study will reveal, nevertheless, more common features, such as the fundamental structure.

### 2.2 Data journals

A first survey on data journals was conducted by Candela et al. (2015), with a sample of 116 data journals published by fifteen different publishers. They distinguished seven “pure” data journals publishing only data papers and 109 “mixed” data journals publishing any typology of paper including data papers. The most represented subjects (in terms of number of journals) were medicine (53%), biochemistry, genomics and molecular biology (26%), and agricultural and biological sciences (16%). They identified only nine data journals in social sciences and humanities (8%). Their results show a recent and slowly developing landscape (the average number of data papers per journal is less than ten), with conceptual, structural and terminological diversity (they identified ten different terms assigned for data papers). Also, there is no consensus about the usual content; the only section present in all data papers is the data availability (location, accessibility), followed by information about the provenance of the dataset. Most of the data journals perform some kind of traditional peer review to guarantee a certain level of the papers’ quality but also to assert some quality of the datasets in terms of utility and reusability; only a few journals adopted an “open peer review.” Most journals are published in open access, with an average APC amount of 1,300 euros.

*The Grey Journal*, published by Textrelease (Amsterdam) is one of those “mixed” data journals. Initially a regular journal with papers from international conferences and original research articles, *The Grey Journal* started to publish a collection of data papers in 2017. This collection was born out of an “Enhanced Publications Project,” fueled by the FAIR Data Principles (Farace et al. 2018). The main pillars for this collection are the International Conference Series on Grey Literature, the research data that is created and archived within this framework (actually thirty-seven datasets housed in the Dutch data repository DANS), and the existence of a flagship journal for the publication of the data papers. A standardized template is provided to ensure the identity and longevity of the collection and to guide prospective authors and researchers in submitting a data paper. The template consists of five sections each of which has a note field providing examples and/or a maximum word count. The fields are labelled as follows: overview, methods, data description, potential reuse, and references. Currently, seven of GreyNet’s thirty-seven datasets are supported by a data paper (19%). Yet, even on a small scale this data paper collection illustrates an operational and functional ecosystem of open science constructed year after year with five main elements, i.e., an academic community, original research within this community, conferences, a journal, and a data repository. In this emerging framework, data papers gain their particular relevance, different from regular articles.

### 2.3 Features and metadata

Yet, other aspects appear challenging the idea of a clear distinction between data papers and regular papers. Li et al. (2019) conducted a content analysis with eighty-two data papers from sixteen journals to investigate what information they describe regarding the methods to create and manipulate the data objects (i.e., “data events”). For Li...
and his colleagues, even if they have distinct features from research articles, data papers are “nevertheless created under similar conditions,” and they reveal “functional overlaps” between both categories related to the narratives of data events (natural language) and to their composition, which is “inevitably situated in the specific epistemic communities.” Their main function is to improve the findability of published datasets and, through enriched metadata description, to foster their reusability.

Metadata are constitutive for data papers. Candela et al. (2015) produced a conceptual map of the data paper (see Figure 1). They insist not to confuse the data paper’s content, its metadata, and the datasets’ metadata. “The concept of data paper has at least two elements that have to be materialized into concrete and identifiable information objects in order to fully implement it: the dataset, i.e., the subject of the data paper, and the data paper itself, i.e., the artefact produced to describe the dataset” (1752).

The link between data papers and the metadata of research data is essential, because both have similar functions (describe data, define accessibility, (re)usability, and content. Insofar data papers are about deposited datasets and insofar deposits require metadata, data papers can be (partly) derived from existing metadata.

Chavan and Penev (2011, 7) describe a tool that “facilitates conversion of a metadata document into a traditional manuscript for submission to a journal” for biodiversity resource datasets. The human contribution is minimal if the metadata is standardized (with controlled vocabulary), exhaustive, and of sufficient quality: “Once the metadata are completed to the best of the author’s ability, a data paper manuscript can be generated automatically from these metadata using the automated tool … The author checks the created manuscript and then submits it for publication in the data paper section through the online submission system of an appropriate … journal” (7).

This kind of generated data paper can be further enhanced in different ways, such as “describing fitness for use of data resources, which will increase the usability, verifiability, and credibility of those resources,” persistent identifiers, an “interpretive analysis of the data (which) could include taxonomic, geospatial or temporal assessment of data and its potential of integration with other types of data resources” or the inclusion of “a taxonomic checklist and/or the data themselves.” Data papers represent a highly standardized type of publication, with a standard structure and a content, which is largely defined in terms of metadata formats (such as DataCite Metadata Schema) and identifiers for datasets, persons, etc. (such as DOI and ORCID).

2.4 Production and processing

In fact, Chavan and Penev (2011) describe an integrated workflow of data repositories and journal platforms, requiring shared standards and formats. Senderov et al. (2016) provide an example of this data paper generation in the field of biodiversity. Their workflow relies on three key standards (RESTful API’s for the web, Darwin Core, and EML) and imports metadata into the ARPHA writing tool (AWT). In other words, and more generally spoken, “the boundary between a workflow tool, a data store, and a publishing platform blurs” (de Waard 2010, 9).

But are data papers produced only for machines? No, according to Li et al. (2019, 18) who are convinced that “as a genre built upon natural languages, data papers are primarily a human-readable document, much less designed for reproducing data workflows in computational approaches.” Both are complimentary rather than competitive.

In her review of data papers, Reymonet (2017) compares data papers and data management plans (DMP). Indeed, as the expected structure of such an article may be based on the items provided when preparing a DMP, Reymonet suggests a tool (or workflow) to export selected items of DMPs in order to prepare or generate a data paper.

A general assumption is that data papers, like regular papers, are peer reviewed, implying some kind of quality control and selection. This means, too, that metadata of

![Figure 1. Data papers concept map (from Candela et al. 2015, 1752).](image-url)
research data (and, indirectly, the datasets themselves) become the object of scientific evaluation, which “contributes to the popularity of data papers in increasingly more scientific fields” (Li et al. 2019, see also Costello et al. 2013). For the same reason, data papers contribute to the trustworthiness of research data. For example, Elsevier’s Chemical Data Collections invites authors to submit data papers, because this “ensures that your data ... is actively peer reviewed.”6 As cited above, Pärtel (2006) mentions that data papers were about “data sets of value in a given field,” which implies a selection by the authors themselves, upstream of the writing of data papers and of peer reviews even if the criteria of selection remain uncertain.

2.5 Critics and outlook

Similar to most cited authors, Smith (2012, 15) states that data papers “are like traditional research papers in some aspects: they are formally accepted, they are peer-reviewed, they are citable entities” but then adds that “in other respects they are very different from traditional research articles because they are not about the research, they are about the data.” And this exactly is the main reason for some more critical voices, expressing concerns about the real demand by society and research, about the additional workload for authors and peer reviewers, and about the motivation of scientists to share their data. The underlying idea is that scientists should (and mostly do) publish about results not about data.

Other arguments against data papers are their price (APCs) and the slow uptake, at least initially. “To address professional recognition and data quality control, there are viable alternatives to the data paper (such as the) implementation of a joint data-publishing and -archiving policy by databases and journals ... instead of popularizing a new kind of publication, it is more important to improve current peer-review processes and the operating policies and integration of journals and databases” (Huang et al. 2013, 5). Huang’s critic may be specific for a given field of research (here, biodiversity) but should be taken into account for a general understanding of the future development of data papers.

Nevertheless, data journals and data papers appear to be here to stay. The French national plan for open science recommends (MESRI 2018, 6-7) “as part of its government support for journals ... the adoption of an open data policy associated with articles and the development of data articles and data journals.” While data papers become a legitimate (mainstream) part of the landscape of academic publishing, only few studies provide empirical or conceptual elements of an answer to the question of how exactly data papers contribute to the organization of scientific knowledge, compared to regular research articles.

3.0 Methodology

In order to analyse specific features of data papers, we established a representative sample of data journals, based on lists from the European FOSTER Plus project,7 the German wiki forschungsdaten.org hosted by the University of Konstanz,8 and two French public research organizations.9 The complete list consists of eighty-two data journals, i.e., journals that publish data papers. They represent less than 0.5% of academic and scholarly journals. For each of these eighty-two data journals, we gathered information about the discipline, the global business model, the publisher, peer reviewing, etc. The analysis is partly based on data from ProQuest’s Ulrichsweb database, enriched and completed by information available on the journals’ home pages.

Some data journals are presented as “pure” data journals stricto sensu, i.e., journals which publish exclusively or mainly data papers. We identified twenty-eight journals of this category (34%). For each journal, we assessed through direct search on the journals’ homepages (information about the journal, author’s guidelines, etc.) the use of identifiers and metadata, the mode of selection and the business model, and we assessed different parameters of the data papers themselves, such as length, structure, linking, etc.

The results of this analysis are compared with other research journals (“mixed” data journals) that publish data papers along with regular research articles in order to identify possible differences between both journal categories on the level of data papers as well as on the level of the regular research papers. Moreover, the results are discussed against concepts of knowledge organization.

4.0 Results

Four of the twenty-eight data journals have ceased, and two have merged. All of them are published online while nine still have a print version. One data journal is a report series.

4.1 Research disciplines

Most data journals are from STEM domains, in particular from life and medical sciences, including genetics (see Figure 2). Only four journals publish data from the humanities (psychology, archaeology) and social sciences. One data journal covers a large range of disciplines from sciences (Scientific Data by Nature), another is open for all topics in social sciences and humanities (Research Data Journal for the Humanities and Social Sciences by Brill).

The five data journals with papers on data from the arts, social sciences, and humanities represent 18% of all “pure” data journals. In terms of articles (see below), they
represent less than 4% of all data papers published in data journals, with estimated 400-450 papers, mostly in archaeology.

4.2 Publishers

Except for Taylor & Francis, all big five academic publishers (Elsevier, Springer-Nature, Wiley-Blackwell, Taylor & Francis, and SAGE) have their own data journals. Five data journals are published by Elsevier (from which two are published by Academic Press, an imprint of Elsevier, two others merged), two by Wiley, one by Springer-Nature, and one by SAGE.

Other data journals are published or hosted by newcomers, especially by open access publishers such as Ubiquity Press (three journals), BioMed Central (two journals) Hindawi, MDPI, Copernicus Publications, Pensoft, or Faculty of 1000, by smaller publishing houses like Brill or de Gruyter (Sciendo), or by learned societies or university presses (AIP, ACS, Wageningen, etc.).

Most of the data journals are published in three countries, i.e., the United Kingdom, the United States, and The Netherlands. The other journals are from Bulgaria, Switzerland, Germany, and Poland (Figure 3). All are published in English; only one journal also publishes papers in another language, Dutch (Research Data Journal for the Humanities and Social Sciences).

4.3 Business models

Most of the data journals are “young” products with a short history. Only seven journals were launched before 2000. The other twenty-one journals have been launched during the last ten years, from 2008 on, and especially in 2013 (seven journals) and 2014 (five journals). Four journals have ceased or are suspended.

At least one part of the data journals is considered as a good or high-quality journal. Eleven data journals are indexed by Clarivate Analytics and eight by Elsevier’s Scopus database. Sixteen journals are referenced in the international Directory of Open Access Journals (DOAJ).

The overall number of data papers published by these data journals is approximately 11,500, with large differences, ranging from some papers up to more than 3,500. The median number, however, is rather low, with ninety-seven (Figure 4).

In terms of volume, Elsevier’s Data in Brief is by far the most important data journal, followed by Elsevier’s “old” Atomic Data and Nuclear Data Tables (launched in 1979) and Scientific Data, a Nature Research journal from Springer Nature. Together, these three journals represent more than half of the data papers published in pure data journals.

The major business model is OA Gold, mostly with APCs (nineteen) but also without (two). Four journals are hybrid, and only one journal is available through the traditional subscription model (Figure 5).

In this small sample, there is no “diamond OA journal” without subscription and APCs. In other words, twenty-five journals (89%) are OA journals or allows OA publishing, and in twenty-three journals (82%) authors have to pay for OA. All data journals covering the arts, social sciences, and humanities are OA journals, all with APCs.
Figure 3. Geographical origin of data journals.

Figure 4. Number of data papers per journal (with best estimates).
4.4 Licensing

Twenty-one data journals disseminate data papers with an open license, most often a CC-BY license, sometimes together with a public domain license (CC0) or the more restrictive CC-BY-NC-ND or CC-BY-NC-SA licenses (no commercial re-use). Elsevier also proposes its own user license. Only one journal does not propose an open license for the dissemination of the data papers but keeps the full copyright (Journal of Physical and Chemical Reference Data).

4.5 Selection

Except for one title (European Power Watch), all data journals explicitly inform about some kind of formal selection procedure. Often the information for authors just mention “peer review,” but six describe the selection as a single-blind review process where the identities of the reviewers are not disclosed to the author(s). One journal (Chemical Data Collections) applies a “quick peer review” with focus on the data value and potential re-use but does not explain who does the peer review and how long it takes.

Five data journals apply some kind of innovative open peer review, either as an option (if required) or for all submitted papers. Yet, this term has different meanings:

- the reviewers are suggested (and known) by authors (F1000Research);
- community peer review (Biodiversity Data Journal);
- interactive public peer review (Earth System Science Data).

The last procedure is particularly interesting: all referee and editor reports, the authors’ response, as well as the different manuscript versions of the peer-review completion (post-discussion review of revised submission) will be published if the paper is accepted.

4.6 Structure and length

We already mentioned that it is generally assumed that data papers are short texts, up to four pages. In fact, this is only partly true. In this sample, only five journals require short papers, limited to four to six pages or maximal 3,000 words. Most journals do not limit the length of submitted papers or make the usual recommendations (six to ten pages, or maximal 6,000 words). One journal only accepts short abstracts (Ecological Archives), while others publish papers well beyond the length of regular papers, up to twenty or thirty or even 100 pages, including detailed data descriptions, illustrations (figures), or data tables like Atomic Data and Nuclear Data Tables. On the other hand, data journals in the field of the arts, social sciences, and humanities publish generally shorter data papers.

No results, no discussion, no conclusion: usually the data journal guidelines for authors contain these or similar recommendations, like Elsevier’s Data in Brief, which asks authors to “avoid using words such as ‘study’, ‘results’, and ‘conclusions.’” Quite different, the Atomic Data and Nuclear Data Tables guidelines leave it to the authors whether or not to include results, discussion, and conclusion to the description of the data.

Nearly all journals require or suggest a particular structure, and some of them provide a template with mandatory sections. However, there is no standard structure. Instead of a generally accepted succession of sections, data papers are made of three constitutive elements, i.e., an in-
roduction with information about the context and the rationale, a more or less detailed description of the datasets with specifications (sometimes formalized as disciplinary or generic metadata of data, such as the DataCite Metadata Schema or the DDI\(^1\)), and a section of materials and methods, instrumentation, on the production of the data and procedures, sometimes extended to experimental designs and calculation (Figure 6).

The figure presents a core structure with three central sections (in blue), with other, optional or peripheral sections, some of them similar to regular papers (in italics), others characteristic for data papers, such as:

- Value and validation: information about the (potential or real) value of the datasets and the quality control (validation), like peer review, automatic procedures (technical validation), etc.
- Potential reuse: information about potential usage, about reuse, and the potential interest for scientists or other users.
- Access and availability: information about the address of datasets (repository, URL) and the availability, including access and reuse rights and limitations; this part may include implementation details, about the availability of source code and requirements, and about the availability of supporting data and materials.

Information about access and availability may also be part of the appendices, like acknowledgements, references, competing interests, author roles and information, rights and permissions, or even peer review comments.

As mentioned above, some data journals allow or invite sections about results of data analysis, together with a discussion of these results and an outlook on further research, very similar to the usual structure of scientific articles and blurring the frontiers between both types of papers.

A last aspect: no invitation or guidelines were found concerning machine-based generation and/or automatic processing of data papers. Apparently, the publishers’ platforms do not support automatic ingestion of text files (via FTP of repository metadata or similar) but require manual deposits of manuscripts and authorship. Of course, this requirement does not exclude partly or complete machine-based generation of data papers upstream of the human deposit of manuscripts.

4.7 Metadata and identifiers

Two types of metadata must be distinguished regarding data papers, i.e., metadata of the described datasets, and metadata of the data papers themselves.

- Metadata of datasets: as mentioned above, some data journals require a detailed and formalized description of datasets in a format that is potentially compliant with metadata. But only a few journals insist on a specific standard. Two examples: *Ecological Archives* expects strict adhesion to the metadata content standards derived from a set of generic metadata descriptors published by the Ecological Society of America (Michener et al. 1997); the metadata set should be sent to the editor as a separate text file. *Genomics Data* requires compliance with an internal standard for data description with eight fields.\(^1\) Both formats have in common that they are community-specific, disciplinary metadata standards. A

![Figure 6. Sections of a data paper.](image)

---

\(^{1}\) For a more detailed discussion of metadata standards and formats, see the authors' previous work on the topic.
third example is quite different, generic, and limited to the datasets’ identifiers: Scientific Data requires an ISA-Tab\textsuperscript{14} metadata text file where the DOI of all datasets are mentioned.

- Metadata of data papers: most journals ask for some general and usual information, compliant with the Dublin Core format, such as author, organisation, title etc. F1000Research recommends XML Schema, Xlink, MathML, or the NLM Journal publishing DTD (JATS\textsuperscript{15}).

Twenty-six journals publish the data papers with a DOI (93%), and five also include the author identifier ORCID (18%). Also, most of them recommend if not require a standard identifier (DOI) or at least a stable address for the described datasets.

4.8 Linking

All data papers provide information about the availability of the described datasets, mostly together with an address (URL), but they do it in different ways:

- usually in a special section of the paper with a statement on data access and availability,
- in an appendix that contains a declaration with data availability and address,
- in the abstract,
- as part of the metadata.

Some papers contain downloadable data; others require that the described datasets should be deposited in one or a shortlist of recommended repositories.

5.0 Discussion

5.1 A new ecosystem

Compared to former studies, the number of data journals and papers appears to increase slowly on a low level. Garcia-Garcia et al. (2015) identified twenty pure data journals; four years later, our sample consists of twenty-eight data journals and not all are still active and even pure (see below). Twenty-eight journals represent less than 0.01% of the academic and scholarly serials (source: Scopus). The arts, social sciences, and humanities are nearly missing (two journals in 2015; four in 2019). The number of data papers progressed at a faster pace, from 846 in 2013 (Candela et al. 2015) to an estimated number of 11,500 data papers in 2019. Yet, this volume represents roughly 0.4% of the overall number of articles published in 2017 (source: Scopus).

Also, the interest of data papers and journals is not their volume but the fact that they clearly are a product of the emerging ecosystem of data-driven open science. Four aspects characterise this embeddedness in the new environment:

- **Business model:** The dominant business model (gold OA with APCs) is different from the traditional and still prevailing serials landscape, and it appears already compliant with the requirements of the new plan S.\textsuperscript{16}
- **Reuse rights:** most data journals allow publishing with an open license, often with generous reuse and remixing rights (e.g., CC-BY license and/or CC0 waiver).
- **Findability:** the editorial model of data journals requires standard identifiers for the datasets, e.g., DataCite’s DOI, to guarantee (and increase) the findability of datasets; they also attribute DOIs to their own data papers, creating a kind of cross-linked DOI system between data papers and datasets.
- **Interconnectedness:** perhaps the most relevant aspect is the integration of data journals and papers in a complex structure of open access journal platforms and data repositories, academic communities, research projects, conferences, etc. Interconnectedness requires interoperability between platforms and infrastructures but is more than technology, formats, and standards, insofar it means new ways of doing science, including research management, research environment, workflows, etc.

A fifth aspect, i.e., evaluation and selection, is already visible but still in transition and not dominant. Data journals replace the usual evaluation and selection procedure (double-blind peer review) by partly open single-blind peer review and, for already one out of five journals, by some kind of open peer review, including innovative community peer review and interactive public peer review. They can also contribute to the assessment of data value through the follow-up of citations (Belter 2014).

5.2 FAIR principles and beyond

Most data journals have never been produced as traditional serials but are a pure (and young) product of this new ecosystem of open access, open (and big) data, and new forms of selection and dissemination. This makes them particularly different from other academic journals. And this makes them also particularly interesting for the requirements of the so-called “FAIR Guiding Principles for scientific data management and stewardship” (Wilkinson et al. 2016). Their data papers contribute to these principles in different ways, in order to improve the findability, accessibility, interoperability, and reuse of research data, e.g.\textsuperscript{17}:

- **Findable**
  - F2. Data are described with rich metadata: data papers enrich existing metadata of datasets.
- F4. (Meta)data are registered or indexed in a searchable resource: the enriched metadata are registered, indexed, and preserved on the data journal platform.

- Accessible
  - A2. Metadata are accessible, even when the data are no longer available: the accessibility of metadata published via data papers does not depend on the datasets’ accessibility in a data repository.

- Interoperable
  - I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation: at least some of the data journals insist on the application of formal, standard language (vocabularies) for the description of datasets. As a minimum, they reproduce the data repositories’ own formal dataset representation.
  - I3. (Meta)data include qualified references to other (meta)data: data papers can (and usually do) provide links to other related resources, e.g., research papers, institutional affiliations, similar or related datasets, etc.

- Reusable
  - R1.1. (Meta)data are released with a clear and accessible data usage license: as mentioned above, most data papers are published with an open license; whenever the data paper is derived from the original metadata, this license may depend on the repository’s initial licensing and reuse rights.
  - R1.2. (Meta)data are associated with detailed provenance: one of the main functions of data papers is to provide detailed knowledge about where the data came from, who to cite, who generated or collected it, and how has it been processed (workflow).

Along with metadata, data papers contribute to the compliance with FAIR principles, in particular to the two principles of findability and reusability, insofar they help people (and machines) find datasets and inform about the provenance and reuse rights. Additionally, data papers contribute to another aspect, beyond the FAIR principles, i.e., the evaluation of the datasets’ quality and value.

In the context of open science, metadata has been considered fuel for economy (Neuroth et al. 2013). As a new vector of communication of metadata on research data, data papers can be defined as a kind of pipeline for this fuel. Yet, as they also add value to metadata, through contextual information, evaluation, new identifiers, etc., they are not only pipelines but also refineries, more or less specialised, more or less standardized. To stay with the fuel metaphor, data papers are a new infrastructure of refinement and dissemination of the metadata fuel.

Regarding knowledge organization, two aspects require attention and further investigation:

**Standardization**: the quality of data papers depends for much on the quality of the metadata of the underlying datasets; and this means, on controlled terminologies, on standard formats, well-defined elements, etc. One example is the International Geo Sample Number (IGSN) designed to provide an unambiguous globally unique persistent identifier (PID) for physical samples (specimens) and to facilitate the location, identification, and citation of physical samples used in research.¹⁹ The development of data papers and data journals should (will) be accompanied by further work on standards, by academic communities, publishers, information professionals, and knowledge practitioners.

**Specialisation**: to be relevant and useful, metadata standards should be as compliant as possible with the specific requirements and features of scientific communities, disciplines, methods, tools, and equipment. This specialisation, however, tends to limit their interoperability between different domains, infrastructures, information systems, and their interest and usefulness for interdisciplinary research, discovery tools, etc. One solution to this problem could be described by “as specific as possible, as generic as necessary,” an approach that would apply a kind of ad-hoc-compromise for each particular situation, resulting in many different formats more or less specific, and more or less generic. Another, perhaps more realistic approach would be to accept (and support) two (or more) different standards for each dataset and each data paper, one generic (like, for instance, the DataCite metadata schema), the other specific, depending on the particular domain, method, tool, etc.

5.3 Blurred boundaries

The specific identity of data papers is mainly defined in opposition with regular research papers (see for instance Penev et al. 2012). The reality is different. The empirical data of our survey provides evidence that despite a general definition of data papers and journals, there is a lot of divergence and heterogeneity that can be described on four levels.

1. Data journals also accept other articles. Our survey put the focus on a limited number of academic and scholarly journals indexed by databases or directories as “pure” data journals. Yet, even in this sample some data journals publish regular research articles, reviews, short communications, or comments along with data papers, such as Data from MDPI and Earth System Science Data from Copernicus.

2. Data papers are published in other journals. As mentioned above, one limitation of our survey is the fo-
4. There are other emerging types of articles, similar to but not identical with data papers. "Pure" and "mixed" data journals are open for other categories of articles that are neither traditional journal items (research articles, reviews, comments, etc.) nor data papers. Sometimes the difference may be a question of terminology. For instance, F1000Research accepts "brief descriptions of scientific datasets that promote the potential reuse of research data and include details of why and how the data were created" called "data notes" \(^{22}\) — in other words, data papers. But there are other examples (see also the listed terms in Candela et al. 2015):

a. Data services paper: "papers on data services, and papers which support and inform data publishing best practices (including) the development of systems, techniques or tools that enable data analysis, data visualisation, data collection and data sharing (and) processes and procedures used in the development of datasets" \(^{(Geoscience Data Journal)}\).

b. Meta or overlay articles: "Descriptions of online simulation, database, and other experiments, partnering with digital repositories on ‘meta articles’ or ‘overlay articles’, which link to and allow visualisation of the data, thereby adding an entirely new dimension to the communication and exchange of data research results and educational materials" \(^{(Data Science Journal)}\). \(^{21}\)

These two examples of a new kind of paper are quite different, yet they have in common that they are both linked to research datasets and, above all, to the dissemination and reuse of research data, which is their main purpose.

The boundaries between data papers and data journals and other categories of scientific communication are partly blurred, not only due to a lack of reference definitions but also due to a large diversity of publishing practices. This may have at least three explanations:

- The publishing of data papers is still in transition. It took some decades to develop and accept the IMRAD format as a standard format of scientific article publishing.\(^{23}\) The heterogeneous character and blurred boundaries of data papers may reflect the emergence of a young and new; still not well-defined form of scientific communication.

- The described proximity with research communities, the "embeddedness" in an ecosystem defined by disciplines, materials, methodologies, tools, etc., contributes to the heterogeneity of data journals and papers. Data papers necessarily depend on the community-specific way of how data is produced, collected, processed, preserved, reused, and it seems quite natural that they will reflect the diversity of this environment. Perhaps, fuzziness is a core element of the data paper category.

- One part of the new OA journals announces an inclusive editorial policy. Instead of a selective approach and guidelines with explicit limitations, they invite submission of all kinds of papers; a strategy somewhere between predatory publishing and big data principles based on volume and variety rather than on quality and trustworthiness.

5.4 Who is writing? Who is reading?

Some of the underlying questions of this study were about the production and use of data papers. How are they written, and are they really "written"? Which is the (potential and real) part of automatic or semi-automatic production, and which is the part of human added value? In fact, are data papers written by machines and to be read by machines?

The answer to these questions is neither yes nor no. As mentioned above, data papers can be at least research data available in data repositories such as Dataverse or others (see the Pensoft workflow, Chavan and Penev 2011). The technology is there. Recently, the French National Institute for Agricultural Research (INRA) updated their Dataverse-based repository including an online tool that partly generated "by machines," i.e., through the exploitation and trans-
formation of metadata on researchers can use to generate data papers from the deposits’ DOI, in an open text format compliant with INRA’s own data journal or with Elsevier’s Data in Brief.  

Both examples, the Pensoft workflow as well as the INRA tool, reveal the potential of automatic generation of data papers but also its requirements and limits. Automatic generation of data papers requires a high degree of standardization and interoperability between data repositories, text processing tools and journal platforms, especially regarding metadata formats and identifiers. Our study was not about metadata formats of data repositories and about their degree of standardization. But our study reveals a lack of standardization on the other side, i.e., the journal platforms. Paradoxically, this may be an opportunity for automatic generation and ingestion of data papers; yet it will not be helpful for machine-based exploitation of data papers.

The limits of automatic generation of data papers are twofold. On the one hand, journal platforms still and always require authorship, i.e., intellectual property and institutional affiliation. They do not accept automatic submission of machine-produced data papers. On the other hand, the format of data papers requires rich contextual information that may not be part of the datasets’ metadata and must be added by the researchers or data officers. Candela et al. (2015) mention that the metadata is usually selected by both the data journal editor (for the data paper) and the data archive manager (for the dataset), which “often results in proprietary, ad-hoc-solutions.” Relevant for our question is the human contribution (selection) and the resulting diversity and specificity.

Candela et al. (2015) also insist on the distinction between metadata of datasets, metadata of data papers, and data papers themselves. Metadata are made for machines, and the main purpose of FAIR principles is to improve machine readability and transfer of research data. Data papers are part of this ecosystem, and they contribute to the automatic processing of research data and related metadata. However, the state of the art and our empirical results (still) reveal human added value, i.e., enhancement of the information produced by metadata, such as potential reuse (value), related datasets and research, and other contextual information useful for the understanding of the described data. But as mentioned above, this can also include more traditional content, like results of data analysis and rich textual discussion of data and results. Another “human added value” is the intellectual responsibility and property of the data papers, which are always attributed to people (authors) not to machines. Instead of machine generated data papers, we should speak of “machine-(or repository-) assisted writing of data papers.”

So, are data papers written for machines? Penev et al. (2012) insist on the “human-readability” even of automatically generated data papers. Rich and less standardized and coded textual discussion, for instance, is probably more aimed at human readers. This of course does not exclude the potential of data papers for automatic exploitation with tools of text and data mining (artificial intelligence). Similar to the generation (writing), this potential depends on the standardization of data papers, including careful coding, and their own metadata, i.e., standardized and well controlled formats and terminology. Probably, the fast development of artificial intelligence will facilitate the automatic production as well as the automatic exploitation of data papers and their metadata. However, so far, we did not identify any study about this potential, which for the moment apparently remains theoretical.

5.5 Data? Information? Knowledge?

Finally, what is the informational status of data papers, compared with the DIKW model of information sciences (Rowley 2007)? What do they carry: data, information, knowledge, or wisdom? Following the usual definitions, the answer seems easy: insofar data papers provide description of data, and insofar information is inferred from data and contained in descriptions (Rowley and Hartley 2008), data papers correspond to the second level of the DIKW pyramid, i.e., information (Figure 7). They are not knowledge but contribute to the generation of knowledge. Also, the main purpose of data papers—to facilitate the findability and the reusability of research data—is similar to another general aspect of information, i.e., its immediate usefulness for decisions or actions.

This characteristic of data papers is one major difference with regular research articles, which are expected to provide more than simple descriptions of facts (data), i.e., insight, understanding, interpretations, hypotheses, etc. However, as mentioned above, the boundaries are partly blurred and some data papers do more than carrying information about data, in particular when they include sec-

![Figure 7. Data papers and the DIKW pyramid.](attachment:image.png)
tions with data analysis results and discussions. So at least partly, data papers also convey knowledge, even if this is not part of their core function.

Downside of the pyramid, the boundary to the data level seems equally blurred. Because, as described above, data papers do not only provide information about data but can be exploited as raw data on their own, generating information about research projects, scientific cooperation, etc. This means that data papers are also partly data.

For both reasons, data papers do not just improve the referencing of datasets on repositories but fulfill other roles. Their particular information profile can be described in terms of library science, as an original integration (or merging) of writing, cataloguing, and indexing, facing major challenges like standards and terminology. Perhaps data papers are a kind of new boundary object (Star and Griesemer 1989) on the front line between academic publishing and research. Our analysis confirms the statement that data papers are like traditional research papers in some aspects but very different in other respects (Smith 2012). Perhaps data papers are not (only) part of academic publishing but should (also) be considered and assessed as part of research data practice.

6.0 Conclusion

Data papers have been defined as scholarly journal publications whose primary purpose is to describe research data (facts about data). Yet, the literature overview shows that there is a lack of a generally accepted reference definition of data papers. Likewise, few conceptual studies and empirical evidence are provided. Also, up to now, the success of data papers appears of minor importance and limited to STM disciplines, primarily in the life sciences.

Our survey provides more insights about the environment of data papers, i.e., disciplines, publishers, and business models, and about their structure, length, formats, metadata, and licensing. Core elements of data papers are the data description and methods and materials; depending on the data journal’s policy, other sections are requested or optional, such as value and validation, potential reuse, access and availability, and even results of data analyses and discussion of results.

The discussion section of this study highlights five major aspects of data papers:

1. Data papers are a product of the emerging ecosystem of data-driven open science. They contribute to the FAIR principles for research data management, in particular findability and reusability, and add in some degree to the evaluation of the quality and value of the data.

2. However, the boundaries with other categories of academic publishing are partly blurred, especially with regular research papers.

3. Data papers are (can be) generated automatically and are potentially machine-readable; yet, the human contribution (still) appears vital in terms of intellectual property and richness of content.

4. Data papers are essentially information, i.e., description of data (as defined by the DIKW model) but also partly contribute to the generation of knowledge and data on its own.

As to the two camps, human generated vs. machine generated, if a data paper is created by a human—whether or not machine aided, one can speak of knowledge organization. However, if the data paper is solely machine generated it is difficult to attribute this to knowledge organization (excluding any reference to artificial intelligence). The latter is more aligned with automated indexing, cataloguing, and the like.

In relation to the DIKW pyramid, data papers appear between the levels of information and knowledge given that for some people they are not knowledge but only contribute to the generation of knowledge.

However, if one looks at the metadata fields that encompass a full blown data paper—such as the explicit roles of the researchers/authors, the research methods applied, the description of the data, its usability as well as its limitations, then one may conclude that the data paper provides a fuller understanding of the data/dataset. In itself, the data paper provides a best practice in knowledge organization—i.e. if not an example of knowledge generation.

Part of the new ecosystem of open and data-driven science, data papers and data journals are an interesting and relevant object for the assessment and understanding of the transition of the former system of academic publishing. This means that the quality and the usefulness of data papers partly depend on external variables, e.g., the metadata standards of data repositories, their trustworthiness in terms of data quality but also long-term preservation (certification), etc. Therefore, as mentioned above, quality control of data papers (i.e., some kind of peer review) always implies some kind of quality control or evaluation of the datasets themselves and their respective repositories.

Based on our empirical results and former studies, we would suggest the following definition of data papers, keeping in mind the transitional and necessarily provisional character of each conceptual attempt:

Data papers are authored, peer reviewed and citable articles in academic or scholarly journals, whose main content is a description of published research datasets, along with contextual information about
the production and the acquisition of the datasets, with the purpose to facilitate the findability, availability, and reuse of research data; they are part of the research data management and crosslinked to data repositories.

This definition may not cover all different variants of data papers but will be helpful for a better understanding of what we called “blurred boundaries” and for further investigation.

At this stage, a couple of questions remain open; in particular, the following topics should be addressed:

– Monitoring: how can the indexing of data papers be improved in order to facilitate their identification and follow-up (search engines, databases, data repositories, journal platforms)?
– Business models: what is the risk of predatory publishing of data journals and data papers? Is it different from predatory publishing of regular research papers?
– Disciplines: are data papers as relevant in the arts, social sciences, and humanities as in life sciences, chemistry, etc.? Should their data papers be published in large and multidisciplinary data journals, together with STM, or should they have their own data journals?
– Ecosystem: more case studies are needed on specific links between research data management, academic publishing, and the production and dissemination of data papers in a given environment and community (equipment, discipline, structure, etc.).
– Evaluation: our study did not assess whether (and how) scholars get credit for publishing data papers. This, however, will be a key factor for the future development of data papers.

Garcia-Garcia et al. (2015) wondered if data journals will remain part of the research ecosystem or not. Perhaps they will not. However, it seems probable that the number of data papers will continue to grow and gain importance, perhaps (probably) not via data journals but via increasing hybridization of research journals and journal platforms, and perhaps even through the merging of journal and data platforms. In any case, on the boundary between research data management and academic publishing, data papers will continue to provide a highly relevant object for library and information science, especially for the further assessment of the development of academic publishing and knowledge organization in the field of scientific research.

Notes


2. Source: data from Dimensions https://www.dimensions.ai/


4. See for instance http://researchdata.cab.unipd.it/122/

5. Article processing charges: the fee authors or their institutions have to pay (after the acceptance of their papers) to some publishers to be published immediately in open access. The amount of APC is varying between publishers and journals; the average amount research institutions pay per article is about 2,000 euros (see OpenAPC https://treemaps.intact-project.org/apcdata/openapc/).

6. Chemical Data Collections, see https://www.elsevier.com/journals/chemical-data-collections/2405-8300/guide-for-authors

7. FOSTER portal, see https://www.fostereopenscience.eu/foster-taxonomy/open-data-journals

8. forschungsdaten.org, see https://www.forschungsdaten.org/


10. See https://www.earth-system-science-data.net/peer-review/interactive_review_process.html


13. These eight fields are: organism/cell line/tissue; sex; sequencer or array type; data format; experimental factors; experimental features; consent; sample source location.

14. ISA tools https://isa-tools.org/


16. The plan S gives preference to immediate open access in 100% OA journals, see https://www.coalition-s.org/

17. The description and numbering of the principles follow the GO FAIR list at https://www.go-fair.org/fair-principles/

18. The FAIR principles have been initially designed for automatic data processing.

19. ISGN http://www.isgn.org/

20. CIRAD, see http://ou-publier.cirad.fr/index.php


22. Data Science Journal, see https://datascience.codata.org/about/

23. IMRAD is a common organizational structure of scientific writing and the usual format of papers on orig-
inal research published as articles in scientific journals, in particular in empirical sciences but also in other disciplines. It stands for “introduction, methods, results and discussion/conclusion.” For more details and references, see https://en.wikipedia.org/wiki/IMRAD

24. INRA, see https://data.inra.fr/datapartage-datapaper-web/ and https://dataverse.org/blog/data-inra

25. Metadata considered in the strict sense of the term, i.e., digital data on other digital data.

**Data availability**

The CSV table of the underlying dataset is available in the Dutch repository DANS, at the following address: https://doi.org/10.17026/dans-zk3-jkyb

**References**


Knowledge Organization from a Social Perspective: Thesauri and the Commitment to Cultural Diversity†

Pablo Gomes* and Maria Guiomar da Cunha Frota**

Universidade Federal de Minas Gerais, Escola de Ciência da Informação Av. Antônio Carlos, 6627, Pampulha, Belo Horizonte, Minas Gerais 31270-901, Brazil,
*<pablogomes.pg@gmail.com>, **<frotaguiomar@yahoo.com.br>


Abstract: Knowledge organization systems can have linguistic and conceptual formations of social oppression and exclusion. It is information science’s role to be vigilant in perpetuating seditious discourses which end up reaffirming offenses, prejudices and humiliations to certain groups of people, especially those labeled as marginalized, that is, who are not part of the dominant group holding social power. In the quest for this diversity, this study reviews the literature of the area on how thesauri can become more inclusive and on the role of semantic warrant, specific to the philosophical, literary and cultural warrant. This research highlights the need to review thesaurus construction models so that they can be more open and inclusive to the cultural diversity of today’s society, formed by social actors who claim their spaces and representations. To this end, guidelines are suggested for the construction of thesauri procedures that allow cultural warrant receptivity.

Received: 30 September 2019; Revised: 28 October 2019, 6 November 2019; Accepted: 15 November 2019

Keywords: cultural diversity, cultural warrant, thesaurus construction, information

† Presented at ISKO-Brazil 2019: Organização do conhecimento responsável: Promovendo sociedades democráticas e inclusivas (Responsible Knowledge Organization: Fostering Democratic and Inclusive Societies), at Umarizal (Belém), Pará, Brazil, September 2-3, 2019.

1.0 Introduction

How does one build a thesaurus for some knowledge areas of human and social sciences? The answer to this question may seem quite simple: following thesaurus construction models. However, other questions emerge as we take a closer look at such models: 1) Do they account for the complexity of areas that change so constantly in terms of themes and issues?; 2) Do they cover and cope with issues that are also the subject of disputes in the political and social field?; and 3) Do these models allow the inclusion of cultural diversity addressed by these sciences? The answer to these questions is that to embrace the complexity surrounding cultural diversity representation requires critical analysis and the adaptation or creation of new models.

Thesauri have long become tools for librarians and information professionals to represent information, dating from the 1950s, thus for almost seventy years. During these years, when thesauri contributed to facilitating both representation and information retrieval, little concern was given to the exclusionary relationships resulting from the absence of a more critical analysis regarding the representation forms of diversity.

If we consider a language, in this case documentary language (DL), we need to understand that the use of certain terms in certain settings is relevant in perpetuating forms of social oppression, understood in the context of this research as offensive, biased and humiliating structures to certain groups of people, especially those identified as minorities such as women, black people, LGBT+ and others.
To cope with the terminological and conceptual part of thesauri, warrants for maintaining semantic quality have been included in thesaurus-building processes. Some have been highlighted, such as the philosophical, literary and use warrants. Others, such as the cultural warrant, have failed to find a place in the practical field on the large-scale. Such warrants are responsible for the content these thesauri will have associated with them, whether for diversity or for maintaining discourses of social exclusion.

By seeking the semantic reduction of a term and concept in a given area of knowledge, considering the specialized language (SL) of the field, thesauri need to be aware of the social interlacing certain linguistic formations may reflect on the collection and user communities. Thus, the adoption of warrants ensures inclusion, as the case of cultural warrant.

It is already seen in the knowledge organization literature that DL such as Universal Decimal Classification (UDC) and Dewey Decimal Classification (DDC) have formations with notations that promote, for example, religious intolerance, as the case presented by Trivelato and Moura (2017) when analyzing the UDC. This research, in addition to many others, leads us to resume the discussion regarding the neutrality of the information action. Thesauri do not deviate from the classification tools pointed out above, as they are simplified representations of the conceptual structure of an area through a hierarchical linguistic structure. In their construction processes, they end up being judged by those who participated in making decisions implying partial interpretations and points of view. Guimarães and Pinho (2007) state that representation systems reflect pre-established standards set out by the creator of a knowledge organization tool. Therefore, it is utopian to speak of neutral information systems, instruments, or actions, as the worldviews of the creators of such elements will be implicitly implicated in the process. Thinking of the transition from neutrality to ethical action is crucial, that is, changing the mindset that there is neutrality in information activity to an ethical information activity concerned with the impact that each action may reflect. They are in line with Barité (2011) by understanding that knowledge organization systems (KOSs) entrenched with the worldviews of one single culture are excluded from cultures other than the one privileged in the construction of the system. Their critique of the construction conducting a KOS, either by the creator's interpretation of the instrument or by the worldview a single culture can offer, can produce social oppression by maintaining and continuing dominant discourses, which may be offensive to certain groups of people.

In this context, this research was guided by the question: how can thesaurus construction procedures be better suited to areas such as the humanities and social sciences through cultural warrant and by expanding cultural diversity? The general objective was to propose guidelines within thesaurus construction models that would open cultural diversity through cultural warrant and be able to cope with the complexity of these sciences.

2.0 Thesaurus and cultural diversity

We have lived moments of social construction when minorities, those left apart from the structure of society, claim their spaces and their voices. Thus, rethinking all social elements is fundamental to give due space to all those who are part and contribute to the construction of society. This process implies the inclusion of cultural diversity in all sectors: in labor, academia, politics, science, information, among others. It is crucial that all instances of information science (IS) studies be concerned with the issue of social inclusion of minorities to be able to achieve cultural diversity. Knowledge and information organization should seek to represent, through its instruments, tools and information products, the various forms of cultural diversity, so that greater identification of users, institutions, information systems or services is achieved.

Thesauri are widely used information organization tools by professionals such as librarians, archivists, museum professionals and others. For a long time, these instruments have been known to facilitate information representation and retrieval. However, advances in ICT have led to greater global cultural interaction at an unimaginable speed for over sixty years, and thesauri have failed to keep up with the expectation of indexing information encompassing global cultural diversity.

One of the elements that eventually caused this lack of representativeness in thesauri was its characteristic of bearing a single voice, that is, a single discourse, usually the scientific-academic one. This characteristic provided what Cintra et al. (2002) and Dodebei (2002) named as “economy,” not in the sense of area of knowledge but regarding an economy of meaning, as it reduces the scope of a term within a context.

However, opening up the possibility of including several voices in the same thesaurus does not mean removing their contextual limitation characteristic. To think about the possibility of the thesauri to accommodate the views and values of more than one culture is in line with the studies of cultural hospitality (Beghtol 2002). This is the ability of a KOS to incorporate more than one worldview into the process of knowledge representation. The author seeks to include the various cultures as a preponderant factor in the fulfillment of an ethical posture as a discussion point in KOS. In this way, cultural hospitality would bring the thesauri a multitude of voices.

Culture “leads us to a complex conceptual universe, made up of innumerable theoretical-ideological strands that
reflect the different views on its conception and existing dimensions” (Boccato and Biscalchin 2014, 39, our translation). In this sense, both thesauri and KOS require sensitivity to the inclusion of viewpoints and values shared by various possible stakeholders for the represented information.

Thesauri, as reality representation systems, cannot be exclusive or discriminatory of the ways in which different people, communities and cultures view a particular apprehension about something or someone. The moment thesauri fail to include perspectives on the same object, they set aside a community of potential users to the information system, firstly because those users cannot communicate with the system and secondly because they are not represented or included in it.

We have witnessed cultural interactions closer and closer, and eventually becoming what Garcia Canclini (2015) names “hybrid cultures.” In our understanding, this hybridization is the interaction process involving different cultures, which ends up leading to cultural merge. This perception of hybrid cultures is necessary for information organization when we consider a diversity of people, from different cultures, accessing, interacting and retrieving information in the same information system.

As for knowledge organization, these principles are appropriate to the activities, but not to all the processes permeating this work today. These processes demand reformulation to adapt to the inclusion of cultural diversity. One possibility is to include warrants other than literary and use ones, the most widely used, to include cultural warrant.

3.0 Warrants in thesaurus construction

Warrants are studied in the context of KOSs, and the specific literature on IS offers large scientific production related to classification systems. However, the application of these warrants is also appropriate for thesaurus study and construction. Tennis (2005) highlights that warrants are the rationalization of the choice of a term or concept to be included or not in a controlled vocabulary, as it will give the necessary limits for inclusion or exclusion based on terminology. Thus, Feinberg (2010, 492) states that a warrant “defines the potential sources and rationale by which a classification designer determines the content of the classification.”

Thesauri are constructed and formed based on decisions that cannot be merely taken by someone or by a group’s wish or personality. Solid foundations for making these decisions are built, which are generally related to the inclusion, exclusion or relocation of terms within the thesaurus. Warrants, therefore, “are important to validate the concepts and terms used to represent a given domain” (Dias 2015, 10, our translation). Following this line, Zeng (apud Carlan 2010, 31, our translation) indicates that “the selection process of terms and tests under the principles of ‘warrants’ are very important to the development of any KOS.”

Warrants are fundamental to KOS constructions and, more precisely, to thesaurus constructions. This is due to decisions about the inclusion or exclusion of terms, which must be rationally based to avoid the construction of an instrument with randomly collected terms that does not represent the EL of an area of knowledge.

Warrants are many, and some authors differ on quantity, but in the context of this research, we consider the literary, philosophical and cultural warrant. The first one, literary warrant, is the most commented one when talking about the construction of DL, as Dias (2015 12, our translation) highlights: “literary warrant is a necessary condition for the construction of classification systems, thesauri and other controlled vocabularies.” Its importance stems from the terminology stored in the specialized literature of an area. Thus, this warrant seeks information sources to validate decision-making and collect terms and concepts that will be part of a KOS. This is a positivist view of a term representativeness for an area of knowledge. However, it is highlighted by the technical standard of the American National Standards Institute/National Information Standards Organization (ANSI/NISO, Z39.19 of 2005) as an indicative element to ensure semantic quality. We can verify this fact in the very concept presented for literary warrant, which is the “Justification for the representation of a concept in an indexing language or for the selection of a preferred term because of its frequent occurrence in the literature.” (National Information Standards Organization 2005, 6). The whole process involved in literary warrant seeks to find the basis for proposing a KOS in the documented titles. As KOSs are often designed for specific areas, the language the expert community uses to point out the most appropriate is sought.

Philosophical warrant, also found as scientific warrant is “consistent with scientific and philosophical consensus and is based on the authority of the academia and research” (Guedes and Moura 2016, 83, our translation). To this end, KOS construction should turn to science, or to scientific practices, for decision making regarding the creation of the system. This is one way of considering and building a KOS, based on greater stability and greater standardization.

With this, it is possible to have a warrant which is close to DL’s goals, as they aim at language standardization within an information system. However, it is noteworthy that DLs seek standardization but not a regularity of language use, since many authors comment on the need for revision and adequacy of DL terms. Bliss (apud Guedes and Moura 2016, 83, our translation) comments that “the importance of the
relationship of principles based on classical structures of thought for the development of a classification system suggests joining practical, logical and philosophical principles in guiding the development of classification schemes.”

Beghtol (1986) indicated that philosophical warrant was pointed out as the most suitable for KOS construction by seeking in science, or in academia, the basis for identifying what has relevance to the system. This is due to the constancy science has on the subjects and themes. Its stability in terms of language would vary less frequently than that used by an area expert or users.

Finally, cultural warrant gained prominence especially with Beghtol’s 1986 article “Semantic Validity: Concepts of Warrant in Bibliographic Classification Systems.” This warrant means that “any kind of knowledge organization and/or representation system may be most convenient and useful for people in a culture only if it is based on the assumptions, values, and predispositions of the same culture” (Guedes and Moura 2016, 13, our translation). The “perspective of cultural warrant is a way of reaching notions and ideas which are difficult to recognize by other semantic parameters in which both user communities can identify and KOS are able to represent abstractions of these ideas” (14). Cultural warrant is, therefore, the means of inserting users’ values into a KOS, which will afterwards be accessed by the users themselves. This process brings the users closer to the information system, as it reflects values and predispositions on some of their assumptions.

Pinho (2006, 64), in line with Beghtol, argues that cultural warrant is a “way to add flexibility to knowledge organization and representation systems to encompass aspects related to cultural diversity and make them represented.” This state is important as each group creates and is part of a specific culture. If KOS can aggregate them, these cultural specificities become, as the author puts it, more flexible as they give users a perspective that is known and shared by them.

In Gracioso’s (2010) understanding, the inclusion of cultural warrant would be a way to reinforce semantic relations within KOSs. Thus, at the end of a KOS creation, there would be a relation of meanings closer to the users’ knowledge. It is somewhat predictable that with these culturally closer relationships with users, information professionals could better understand the relationships of meaning at the time of their work with information. Another point Beghtol (2002) makes, which is recognized by Guedes (2016), is the understanding of cultural warrant as an umbrella concept that would hold all other warrants. They are a form of complement supporting cultural warrant. This umbrella metaphor was given Beghtol as early as 1986 and has come back with Guedes (2016) and Guedes and Moura (2016).

Finally, Beghtol (1986, 121) indicates that: Cultural warrants influence the underlying operational rationale upon which classification systems depend for meaningfulness and utility. To investigate cultural warrant beyond the intuitive or observational level, the techniques and findings of such fields as sociology, the sociology of knowledge and social/cultural anthropology would have to be applied to the study of bibliographic classification systems.

We understand that, to be able to include this warrant, addressing sociological techniques that cope with the cultural part to be inserted into the KOS would be necessary. Some to be indicated are those intended for data collection, interview, focus group, observation and could be used to understand the process of signification of the elements to be included in the system, to make it more diverse in terms of worldview.

4.0 Methodological procedures

The methodological strategies were systematized to allow the analysis of thesaurus construction procedures and to allow indications that make it more open and receptive to cultural diversity. Initially, we sought thesaurus construction models that could support the analysis. Among the several found, the Integrated Methodological Model for Thesaurus Construction (IMMTC), in annex I, was the most appropriate in the context of this empirical research. This model was proposed by Cervantes (2009) in her doctoral thesis in IS. The purpose of her study was to systematize the various models, norms and authors to create a model integrated in steps and procedural instructions. The model structured and compiled the construction steps in one single instrument and thus facilitated and reduced research efforts. However, each step proposed in the IMMTC was decomposed to understand each of the indicated instructions in a more complex way. With this, there was a return to the sources used by the researcher.

Next, the critical reading of the model and the theoretical construct on cultural warrant was performed for a comparison and association that suited the guidelines given by construction models based on philosophical and literary warrant. Finally, the collected guidelines were categorized to facilitate the inclusion of proposals that would allow the model to be open to cultural warrant.

5.0 Analysis and discussion of results

After the initial analysis of the model, four founding elements were identified and categorized in the process to favor understanding the data obtained from the reading of the model and its guidelines. The perceived categories were: a) people; b) materials; c) methods; and, d) processes.
This step was followed by the inclusion of elements that allowed the thesaurus to have cultural hospitality. With this, in each of the categories, strategies were conceived to allow the inclusion of more than one voice in the final thesaurus, that is, more voices in addition to science and academia perspectives, which were granted by the philosophical warrant as the holders of EL of an area.

The intention was to open the possibility of having the participation of other social actors that contribute, in the daily dynamics, to the construction of the EL of a certain area. Limiting science and academia as legitimate holders of this language would reduce the importance of other actors and open gaps that could lead to social oppression and to the perpetuation of discourses that are already settled, but in need to be reviewed. It is based on these four elements that the analysis of the obtained data was made.

5.1 People

By analyzing IMMTC, two groups of people involved in the construction of thethesaurus are verified. The first group consists of information professionals who apply the model and manage its execution. The second group are “consultants” to assure the semantic quality to the process, required by an instrument that conceptually represents an area of knowledge.

The two groups were divided and presented in separate groups, as each group has different relations with the construction process and is not to be confused. The formation of the first group is independent of the second one, but the opposite is not true, as the second group depends on the choices made by the first one. Differentiating people into groups was important to make specific indications for each one as they are not general guidelines and would cover all involved.

The group of professionals involved in the construction of thethesaurus used guidelines aimed at the formation of a multidisciplinary team that has the viewpoint and critical positioning directed to the worldview thethesaurus in construction seeks to hold. The issue involved at this moment is the ethical posture of such professionals in the pursuit of greater cultural inclusion. Knowing that the work to be done reflects on how the information system user community will identify.

As for the second group, formed by consultants, the guidance given by the author of the model based on the instruments she uses is that experts be sought. There is a gap on the understanding of who these experts are, but, assuming that the basis of the instruments the author used focuses on philosophical warrant, we understand that the indication of the expert is one with an academic background.

It is noteworthy that this group is of great importance on the final result of thethesaurus. In the guidelines about the role of this group, their main role was to advise about the composition of the bibliographic materials to be consulted for terminological research and for guidance on the conceptual structure of the area. This matters as the worldviews of the first group must be aligned to those belonging to the second group. These understandings should be based on the pursuit of diversity and tolerance.

To bring diversity to the composition of this second group, the inclusion of people outside academia and science is also suggested. It does not mean excluding the participation of this collaborator profile, but the inclusion of other people, such as from social movements, from the technical practice of the area of knowledge, people who work in institutions related to thesaurus and others.

With a more diverse formation of the consultant team, the exchange of knowledge, terminology, perspectives and points of view is possible, which is fundamental for designing thethesaurus hospitable to cultural diversity. The multiplicity of voices in the construction of the instrument does not imply the inclusion of several terms for the same concept, or several concepts for a term, but the possibility of a diverse team to collaboratively reach consensus and assist in decision making.

The choices regarding the formation of the teams are relevant to how receptive they are to differences and diversity. Depending on the management of the teams, only sedimentary, oppressive and prejudiced forms may be conducted; however, on the other hand, it may also bring openness to cultural diversity, characteristic of today’s society.

5.2 Materials

In the traditional models of thethesaurus construction, the terminology of a certain area is recorded on some support: the materials. More specifically, through the guidelines of the analyzed model and the others that supported its elaboration, those are bibliographic materials.

At this point, the guidelines given by philosophical warrant are reviewed, but with the addition of literary warrant. The combination of these two warrants provides the guidance that the materials to be used must be those created within the academia or by science, philosophical warrant, and that the support and informational format would be the bibliographic sources (books, specialized dictionaries, manuals, journals, congress proceedings and others), due to the literary warrant.

In the guidelines of the bibliographic materials produced by science and academia, several of the instructions present in thethesaurus construction models are highlighted. It is understood that the EL of an area would emerge and be created by the specialists, again, those in academia, who regularly search in these materials for specialized terminology.
However, this line of thought, especially in the human and social sciences, presents certain complications as it is excluded from the complexity of an area, the social actors and their production of information and knowledge, which contribute to the constitution, organization and maintenance of discussions that may not have been created by those with an academic background, who hold the title of specialist but who maintain communication language that is not so distinct from so many others. This means there is specialized communication outside science and academia, which is often set aside for the creation of thesaurus and other DLs.

As a way of breaking this maintenance of science as an EL holder of an area in the thesaurus construction process, the inclusion of diverse sources is indicated. Possible examples are to understand and collect the terminology of non-academic books, newspapers and magazines, social movement bulletins, interviews, documentaries, testimonials, etc.

It is also necessary to think about the disruption of using written materials. Audiovisual sources can be important sources for collecting the terminology of certain areas, as the case of the Transitional Justice in Brazil, a branch that has collected testimonies of victims of the Brazilian military dictatorship and which incorporates the specific worldviews and terminology in its reports but to some extent are also shared by science.

The need to reaffirm that the guidelines do not exclude or invalidate the importance of science and academia for the construction of thesauri is again noteworthy, but we highlight the inclusion of new actors and materials created by them in their movements, in their struggles and in practices that involve the development of an area of knowledge.

5.3 Methods

Methods comprise the systematization for thesaurus construction. It is through this set of guidelines, which also represent the team’s choices, that will determine how close or far the final instrument will be to a quality representation of the designed area. The main methods indicated in the IMMTC and the sources are documentary analysis and technical reading of the works. The second one is part of the procedure indicated in the first one, that is, the technical reading is a part of the procedures that comprise what the documentary analysis is. To find the most relevant concepts and terms within the documentary corpus, choosing a technique that allows their identification without having to read all the documents in their entirety is necessary. Thus, documentary analysis emerges, which is widely used by information professionals and studied by information science.

Documentary analysis process begins with reading the document until it is reduced to “products” that facilitate information retrieval and dissemination. One of these products is the representative terms of concepts present in the consulted documents. This implies that these collected terms will be used for the conceptual construction of the chosen area for the thesaurus. Several authors establish the need for understanding the structure of the text as the objective for a good reading, therefore, knowing the macrostructures of the text. This implies using documents with standardized structures that allow this reading with predefined directions. To provide guidelines for documentary analysis, NBR 12.676/1992 was created. It also includes reading the works as a way to identify relevant terms and concepts. However, a relationship of standardization of procedures is perceived.

The new proposal for thesaurus construction methods is to use more open procedures. Considering Bechtol’s (1986) proposal to use research techniques from sociology, sociology, knowledge and social anthropology, the author indicated the use of techniques used to collect data from research in these areas to achieve an approach and an openness to cultural warrant. We understand, therefore, that to include this warrant, addressing sociological techniques that cope with the cultural part to be inserted into the KOS is necessary. Some techniques to be indicated are: interviews, focus group, observation, which could be used to understand the process of signification of the elements to be included in the system, to make it more culturally diverse.

These data collection instruments would be indicated for collecting the opinion of consultants, those who have knowledge, not only scientific and academic knowledge, but also of the analyzed area. It is a way of bringing these consultants closer to the thesaurus construction team. As they include data collection techniques, in this case, especially opinions, the indication of techniques to understand what was collected is necessary. Part of the collected work is easy to understand, but the other part, which is not so explicit, needs methods that help the team understand what the consultants said. The use of discourse analysis and content analysis can be effective at this phase.

Following the indications of documents to be used for terminological collection, the next step is to analyze these documents to verify that their understanding of the world aligns with the view the thesaurus will reflect. Again, the use of discourse analysis and content analysis techniques are important. As books, articles, videos and other documents culturally reflect worldviews of those who have written them, the works need to be aligned with the instrument to be built. As going through the works to collect fundamental terms and concepts is necessary, reading continues as a procedure. However, the use of a less standardized reading is indicated, as it takes into consideration more diverse materials than those proposed in traditional models.
Therefore, the creation of reading strategies for each type of terminological source is needed. The bibliographic strategies are the most common that have systematizations in the information science literature. However, the most diverse sources, such as testimonials, documentaries, personal documents and administrative documents, need more open reading strategies, as they fail to follow standardization. Thus, the strategies would be predefined, but not closed, paths the professional would take to identify relevant terms and concepts for insertion into the thesaurus.

5.4 Processes

The processes in the analyzed model seem to refer to the most standardized forms possible. This must have been a characteristic that allowed great expansion of thesaurus construction, because when the complexity of construction of these instruments is restricted to standardized procedures, they facilitate the construction process. Standardized processes assist in the creation of automated processes.

For a shift toward cultural warrant, the process needs to consider the specificities of each area. Some areas have characteristics that prevent using models as standardized as IM-MTC. Areas still under construction that are intertwined with struggles and social disputes in the material and symbolic fields are a good example, as the terminology of the area is difficult to understand. Disputes involve the process in complex decisions that in some ways represent inclusions and exclusions. Therefore, it is always necessary to turn to forms of inclusion and to an ethical perspective guided by diverse worldviews that fit the thesaurus’s goal.

The process, with all these transformations, require professionals involved in construction who have the knowledge and skills to work with the indicated terminology sources, as well as the methods of collecting opinions, reading literature and other works. This is all based on a critical view that allows constant questioning about the role the thesaurus will play not only in the information organization but also how it will identify with the user community.

6.0 Conclusion

In the course of this research we learned that cultural warrant supported the participation of diverse people, terminological materials, methods and processes for the construction of thesaurus. Sources, people and materials, traditionally focused on the scientific and academic point of view, have been expanded to the plurality of sources also recognized by information science. This leads the thesaurus to the possibility of living a diversity of points of view that have been neglected in traditional models.

The objective of the research was not to disqualify or criticize the model proposed by Cervantes (2009) but to understand and argue that there are knowledge areas and domains that do not respond to standardized ways in their conceptual and terminological construction. This may not be the case in areas such as exact sciences and biological sciences, which tend to have a greater constancy of conceptual and terminological standards.

Extending this model to areas such as the humanities and social sciences is necessary to the point that they do not respond so clearly in their conceptual and terminological construction. Various actors in the field, not only academia and science are involved. Social movements, institutions and professionals often constitute new terminological forms that end up affecting the understanding of inclusion and exclusion.

In times of revision of concepts and terms that may represent oppressive forms of power and at the same time generate social exclusion, it is necessary to think of instruments that include forms of incorporating people and terminological materials in their procedures that allow the representation of an area in an inclusive manner and with a different cultural perspective.

For this, thinking beyond the simplistic view of thesauri being instruments of semantic reduction within the context of information systems is needed. Information professionals must be aware of the role this reduction plays within the collection, and for their user community they can seek greater diversity and inclusion of concepts and terms. This implies making thesauri more democratic instruments of social visibility.

This research does not yet represent a conceptual breakthrough of traditional forms of thesaurus construction, but in its essence, it raises a need for a review of thesaurus models, rules, norms and guidelines. All this to present more dynamic and open guidance about the understanding that not all areas and domains respond equally to conceptual and terminological construction.

References


Appendix I: Systematization of thesaurus construction steps

INTEGRATED METHODOLOGICAL MODEL FOR THESAURUS CONSTRUCTION

<table>
<thead>
<tr>
<th>Systematization of thesaurus construction steps (standardization, literature and thesaurus) – Terminographic procedures</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1. Preliminary work</strong> (General guidance /Use of automated equipment of data processing)</td>
</tr>
<tr>
<td>- Choice of the thesaurus domain and language;</td>
</tr>
<tr>
<td>- Subdomain delimitation;</td>
</tr>
<tr>
<td>- Establishment of the limits of the thematic terminologic research;</td>
</tr>
<tr>
<td>- Consult with domain/subdomain expert.</td>
</tr>
<tr>
<td><strong>2. Compilation Method</strong> (Compilation Approach)</td>
</tr>
<tr>
<td>- Collection of the terminologic corpus;</td>
</tr>
<tr>
<td>- Establishment of domain tree;</td>
</tr>
<tr>
<td>- Expansion of representation of chosen domain.</td>
</tr>
<tr>
<td><strong>3. Term record</strong></td>
</tr>
<tr>
<td>- Term collection and classification.</td>
</tr>
<tr>
<td><strong>4. Term Verification</strong> (Admission and exclusion of terms / Specificity)</td>
</tr>
<tr>
<td>- verification, classification and confirmation of terms;</td>
</tr>
<tr>
<td>- creation of definitions;</td>
</tr>
<tr>
<td>- use of specialized vocabulary to establish relationships among descriptors and relationships among descriptors and non-descriptors.</td>
</tr>
<tr>
<td>- Organization of relations among descriptors.</td>
</tr>
<tr>
<td><strong>5. Thesaurus presentation forms</strong></td>
</tr>
<tr>
<td>- Thesaurus presentation works.</td>
</tr>
</tbody>
</table>

Source: Cervantes (2009, 163, our translation).
Towards a General Conception of Warrants:
First Notes*

Mario Barité

Universidad of the Republic, Uruguay, San Salvador 1944, CP 11200,
Montevideo, Uruguay

Mario Barité is Full Professor at the Faculty of Information and Comunication, University of the Republic, Uruguay. He holds a doctorate and has a master’s degree in scientific information from the University of Granada, Spain. His main research interests include knowledge organization, terminology and e-governance. He is a level I researcher of the Uruguayan National System of Researchers. Among another contributions, he has published a Dictionary of Knowledge Organization. He is editor-in-chief of Informatio Journal.


Abstract: The areas of knowledge are organized around the identification of their terms of reference and the relationships established between them. This is the rational basis of -among others- the methodology for the development of knowledge organization systems. The authority from which to select, evaluate or revise the terminology of these systems is established in relation to any of the twenty-one warrants (literary, cultural, etc.) that have been proposed and studied unequally and autonomously in the literature of the area. This paper intends to introduce initial notes and comments to advance towards an overall conception of the warrant notion. For this purpose, the expression “warrant” is studied as a word of the general language as well as a term of specialized languages. Then, the scope of application of the warrants is established. Next, each warrant is placed in one of the approaches proposed by Hjørland to categorize theories and methods (empiricism, rationalism, historicism and pragmatism). From the above, some lines of research problems are identified. A typological table that includes data on all the warrants established until now is proposed, and the first conclusions are drawn.

Received: 30 September 2019; Revised: 25 November 2019; Accepted: 25 November 2019

Keywords: warrant, warrants, knowledge organization, terminology, classification

* Presented at ISKO-Brazil 2019: Organização do conhecimento responsável: Promovendo sociedades democráticas e inclusivas (Responsible Knowledge Organization: Fostering Democratic and Inclusive Societies), at Umarizal (Belém), Pará, Brazil, September 2-3, 2019.

1.0 Introduction

Terms constitute the basis of the conceptual structures constructed in each discipline or subject field to present in an organized way its main concepts, as well as its premises, principles, theories, categories of analysis, agreements and divergences (Cabrè 1993; Burke 2002, 111-152). The areas of knowledge are organized, reviewed and updated not only based on the identification of their terms of reference but also on the relationships established between them. This is the rational basis of the methodology of vocabulary control operations and, therefore, of the procedures for developing knowledge organization systems, such as thesauri, taxonomies and classification systems (Iyer 2012). The warrants constitute the invisible support, the disguised criteria for the selection of terms for inclusion in knowledge organization systems and information systems in general with the main objective of favoring subject retrieval.

According to a definition given by Beghtol (1986, 110), which can already be considered canonical:

warrant of a classification system can be thought of as the authority a classificationist invokes first to justify and subsequently to verify decisions about what classes/concepts to include in the system, in what order classes/concepts should appear in the schedules, what unit classes/concepts are divided into, how far subdivision should proceed, how much and where synthesis is available, whether citation orders are static or variable and similar questions.

The abovementioned is the first definition of warrant recognized in the literature of the area, so it is possible to speak of a late identification of the concept, especially if it is considered that Hulme had already established in 1911 the initial explanation of what he called “literary warrant,” one of the specific varieties of warrants. It turns out then that the spe-
The mere statement of such an ambitious goal as the development of a general conception may seem excessive or disproportionate. However, there are enough propositions, ideas and contributions of a theoretical and methodological nature scattered here and there in the literature of knowledge organization to begin to assemble the pieces of the puzzle. These brief initial notes, of a generic but comprehensive nature, constitute the first result of exploratory research to be continued.

2.0 Analysis of the notion of warrant

2.1 Warrant as a word of the general language and as a term of specialized languages

A first issue to be noted from the semantic point of view is that “warrant” is a word of the general language as well as a term registered in the language of different specialized fields. Thus, it appears in the terminology of domains such as banking, finance, business, constitutional, civil and commercial law. The term “warrant” has also been used in academic texts as the justification for an argument, the evidence that supports an idea (for example, Nunns, Peace and Witten 2015) or the philosophical discussion of beliefs (Plantinga 1993).

It is not the object of this work to determine whether the word went from the general language to the specialized language or vice versa, although it can be presumed that the need to strengthen commitments, debts and obligations between people, and later between people and/or institutions or states, has accompanied the parallel evolution of the word in the general language, and the term in the mentioned specialized areas. In its semantic evolution, the expansion to meanings such as “foundation,” “justification” and “reason” is also included.

In the territory of knowledge organization, the warrant strengthens in the first place the relevance and adequacy of a term to represent a concept, so that this term can be used for the classification or indexing of data, documents or information. That said, it would seem that the issue of warrants is settled term by term. However, if the warrant will constitute an intellectual criterion to select the term of a knowledge organization system (Huvila 2006), which aims to obtain a consistent and updated subject representation of a domain, then a comprehensive view should prevail. Bullard points out (2017, 76) that the warrant is an element of all classification design, regardless of whether it is named as such and regardless of the particular technological basis of the system. Indeed, warrant is a common thread across a wide variety of systems ranging from traditional library classification to in-application menus and categories for web-based collections because all designers of textual organizing schemas must look to some source of terminology.

From this broad perspective, the warrant should be seen as a tool not only for conventional resources such as a thesaurus, but also for the choice of expressions intended for any grouping of documents or data on the internet. Going a little further, their use and potential could be extrapolated to other areas, such as the selection of the terms to
be included in a specialized dictionary, based on the same arguments and methods.

In order to better justify the need for an integrated epistemological perspective, Olson (2004) tracked the bottom line of the notion of warrant in Francis Bacon's knowledge classification work and its reflections on library classification systems. In that work, she points out that Bacon created a knowledge structure based on a common episteme on which the classificationists have been assembling their system designs from the nineteenth century onwards. And she concludes (Olson 2004, 4) that: “perhaps we should also follow his epistemological warrant and let our classifications not only reflect knowledge, but also have a role in directing the creation of new knowledge.” This neo-positivist perception may have led, in a more or less intuitive way, the classificationists to fuel—over a century and a half—an allegedly common epistemological vision, in order to validate the idea that the precepts of science are objective, neutral and universal in scope.

In a more hidden way perhaps, it was gradually established that the justification of the terminology of a knowledge organization system should be taken in full, either from the formal language of science and science education (Bliss 1929) or from the terminological evidence provided by the documentation (Hulme 1911). This traditional scheme gradually broke down with the proposal of new warrants (cultural, user, organizational, among others), which constituted breakpoints in two ways: 1) in relation to the alleged objectivity and neutrality of the language of science; and, 2) regarding the desirability of maintaining a single warrant for all the terminology selected in the process of developing a knowledge organization system.

When Beghtol states that the warrant is the authority that a classificationist invokes (Beghtol 1986, 110), she places in that authority the source of legitimacy and the ultimate basis for decision-making regarding the terminology to be included in or excluded from a knowledge organization system. She says it unequivocally: the warrant is the authority. This principle of authority may be firmly established in the terminology chosen by those responsible for a knowledge organization system, or it may be more or less blurred, as a consequence of the selection of ambiguous, generic or hybrid terms, or not sufficiently representative, according to the greater or lesser degree of methodological rigor in the selection and systematization of the terminology. A more or less corresponding relationship should then be presumed between the strength and the adequacy of the authority or warrant and the quality of the final terminology product.

Two of the six definitions established by the Dictionary of the Real Academia Española for the entry “authority” are applicable in this context. The third meaning says “prestige and credit attributed to a person and/or institution for its legitimacy or its quality and competence in some matter.” And the sixth meaning establishes that authority is every “text, expression or set of expressions of a book or writing, which are cited or alleged in support of what is said” (Real Academia Española 2014, 246). Although the two meanings seem to fit better with the traditional vision of warrants, considering the social respect for the word of scientists and thinkers, and the veneration for the written word, the truth is that they also allow new authority figures to fit into them: the leaders of certain cultures or minority groups, the alternative terminologies proposed by thinkers or social movements—especially the countercultural ones—or even by the suppliers of new technologies. The doors can only be opened to new authority figures if new warrants different from the traditional ones are introduced, warrants which meet the needs of subject representation of groups, cultures or subcultures of low visibility.

2.2 Scope of application of the warrants

Beghtol’s groundbreaking definition (1986), besides being extremely detailed, is very accurate. For example, it is a merit of hers to propose the incidence of the warrant at two levels or times: that of the initial justification (that is, when creating or selecting a descriptor or subject heading) and in the verification (for example, at the time of the evaluation of the quality and relevance of a term). This implies that the warrant can be used as guidance and as a tool in the three most important processes that a knowledge organization system can go through: its construction, its evaluation and its revision. Therefore, the tuning of specific methodological devices to choose, assess and replace or update their terminology fits here.

Another of the high points of Beghtol’s work is the statement—never contested in the literature of the area—that the warrant must be applied at all stages of the design of a knowledge organization system, namely:

i) in the selection of the terms of classification and indexing;
ii) in the selection of the relationships established between them, an issue also mentioned later by other authors (Rowley 1987; Barité et al 2015, 77);
iii) in the arrangement of the terms of the facets, thus resuming the original application of the literary warrant proposed by Ranganathan in his Prolegomena (Ranganathan 1967, 196);
iv) in the choice of criteria for the subdivision of matters;
v) in the determination of the specificity levels;
vi) in the application of synthesis mechanisms (as in the choice of auxiliary tables or the signs of combination of issues), the selection of syntax devices; and
vii) in the citation order of matters (Beghtol 1986, 110)
At this point, it should be noted that warrants can become epistemological references, organizational criteria or tools at the service of a terminology selection method.

It is noted that the type of warrant applied to the selection of classification or indexing terms may affect not only the decision to include or exclude terms but also the distinction between authorized and unauthorized terms. Indeed, the decision that is taken, for example, to establish in a thesaurus that the term “small states” will be a descriptor, while its synonym “microstates” will be considered a non-descriptor, must be based on a criterion supported by a warrant. The way to warrant relationships between terms has been an issue just outlined in the literature, both from a theoretical and methodological point of view, so it constitutes an open line for research. In these notes, as a first statement subject to review, it is suggested that the warrant procedure for hierarchical relationships between terms should be different from the procedure for associative relationships.

Indeed, the hierarchical relationships established in a knowledge organization system are usually those formally stated and accepted in different disciplines. They are stable, proven and not casual relationships. On the other hand, the systems usually provide the possibility for the classifier, the indexer or the end user to establish an associative relationship between two issues, perhaps because it may be temporary or provisional. Of course, this division of qualities between hierarchical and associative relationships is not absolute, at least as regards a large set of associative relationships that are firmly established in the world of knowledge and documentation (for example, the relationships between church and state, or the relationships between certain plant or animal substances and the treatment of diseases). If this provisional starting point is accepted, the warrant for hierarchical relationships that are established in a knowledge organization system should more reasonably come from the formal classifications of the disciplines (what is known as academic warrant). While to justify an associative relationship the most appropriate warrant for each case should be identified.

2.3 Warrants and epistemological theories

In different works, Hjørland (2003, 2006, 2013, among others) has given evidence of the importance of associating epistemological approaches to the analysis of theories and methodologies in knowledge organization. The most evident advantage of this procedure is that each theory or set of ideas and each method can be inserted into a more general category of analysis, thus taking advantage of all the accumulated flow of reflection made from each epistemological approach. Thus, information science (in particular knowledge organization) is linked in an approachable way to the more general foundations of the sciences and disciplines. Besides, these intellectual tools can be applied to any topic of interest as an object of study for knowledge organization (as is the case with warrants).

The four epistemological-based approaches proposed by Hjørland are: 1) the empiricism, which is justified by the data coming from the set of observations and their corresponding inductions; 2) the rationalism, which proposes the development of knowledge based on principles of pure logic or pure reason and which relies substantially on deductive processes; 3) the historicism, which promotes the organization of knowledge based on chronological, evolutionary and/or contextual studies of each field of knowledge; and finally, 4) the pragmatism, in whose essence the analysis of reality is based on the determination of values, goals and consequences (Hjørland 2003, 2006, 2013). In this exploratory work—in the table presented in Appendix A—each warrant will be related to at least one of the four approaches mentioned on account of future particular analyses.

2.4. Warrant types

In a recent review of the existing literature on the topic, a total of twenty-one warrants proposed, named and used in knowledge organization and close subject fields (Barité 2018, 528) were registered. In that work, there is a synthetic table with the relation of all the warrants. Not all the authors explicitly coined the term “warrant” although it was clear in the respective texts that, in essence, they were talking about the authority invoked to represent knowledge through descriptors, headings, keywords, classification numbers, taxa or others symbols. For the purposes of this paper, the aforementioned table (which due to its length is presented—as we said—in Appendix A), was revised and expanded, with the aim of presenting the twenty-one warrants in a single graphic expression, clear and exhaustive, with the following data: name in English, name in Spanish, author who proposed it and the year of coinage. A column of comments was also added as well as another column that provisionally places each warrant in one of the four epistemological-based approaches outlined above. In the table, the warrants are arranged in chronological order of proposal.

Since the concept of literary warrant (Hulme 1911) was introduced, of the twenty remaining warrants, those most frequently referenced are cultural warrant (Lee 1976; Beghtol 2002a), academic warrant (Bliss 1929; Sachs and Smiraglia 2004) and user warrant (Lancaster 1977; Hjørland 2013). Basically, they are distinguished, because they invoke different sources of authority to collect terms: the language of communities with their own cultural or local identity (cultural warrant), the formal vocabulary of disci-
utilities that each warrant can offer are usually found. This is the case of cultural warrant, user warrant, organizational warrant, academic warrant and more recently indigenous warrant. A third group brings together the warrants that have been proposed and/or mentioned occasionally and that do not have significant subsequent development, as in the case of market, structural or autopoeitic warrants. These last two groups of warrants do not have enough critical analysis yet, and they need it.

From the methodological and applicative point of view, there is a wide research scope on a series of issues that have no definitive answer, and in some cases, not even partial hypotheses or interpretations. These questions can be grouped into two categories: those related to the application of a warrant autonomously and those that propose the convenience of combining two or more warrants to support the box of terms of a knowledge organization system or an information system.

Among the first, there appear the following questions: 1) How do we decide the most appropriate warrant for each system?; 2) How do we guide the choice of one warrant and not another one?; 3) What methodologies does each warrant offer for its application?; 4) In what thematic, documentary or information contexts can a warrant be applied?; 5) How is the “performance” of a warrant evaluated on the basis of the principles of consistency, exhaustiveness, thematic adequacy, linguistic adequacy and other indicators that can be proposed?; 6) How do technological advances contribute to or hinder the application of warrants based on algorithmic mechanisms of automatic or semi-automatic subject assignment?; 7) What theoretical principles and methodologies related to warrants are valid in digital information environments?; 8) How are warrants linked to natural language indexing?; and, 9) On the other hand, can it be proven that the selection of a single warrant ensures the terminological consistency required for the system to be useful for users? In relation to the latter, it is pertinent to wonder, given the variety of warrants that have been proposed by different authors for over a century, if they can be combined or complemented to obtain a better quality and adequacy of the terminology, or if some exclude others, and in the latter case, which do and which do not and why.

Opinions in the literature of the area are divided. Svanberg believes in the complementary use of warrants but also that some may be opposed to others, without providing further explanations (Svanberg 1996). Bullard, on the other hand, expresses (2017, 77) that “the various warrants available to classification designers represent contradictory positions in classification theory yet they compete and are combined by classification designers in daily practice.” After taking a position on the literary, academic, user and ethical warrants, and discussing the possible compatibilities and incompatibilities between them, she states (2017, 77) that “inevita-
ble compromises of daily classification work” [require] “the interaction between warrants.”

Huvila believes that two or more warrants may be opposed to each other, but he proposes to incorporate the concept of hospitality, as reinterpreted by Beghtol (2002a), because (Huvila 2006, 60) it “may be used to denote an ability to incorporate both intra and inter warrant differences i.e. eventual changes within and between individual warrants.” Wan-Chen Lee, on the other hand, generates ideas to understand the nature of the conflict between several semantic warrants, and offers some negotiation alternatives for their use and combination, within the framework of the evaluation processes of knowledge organization systems (Lee 2017).

From this brief review, there appear to be many blind spots that can be used for research on the theory, methods and applications of warrants, in an area in which the emergence of new types of knowledge organization systems has been constant in the last twenty-five years (ontologies, web taxonomies, folksonomies and social classifications, among others), as well as technological innovations that have had an impact on customs and habits in relation to the search of and access to information.

3.0 Conclusions

Warrants are currently seen as an essential component in the process of construction, evaluation and revision of knowledge organization systems, to the extent that their proper understanding and application should ensure consistent terminology, updated and adjusted to the purposes of system designers, and users’ information needs. Moreover, warrants can be used as tools to guide natural language indexing, contributing to correct the undisciplined tagging of social classifications, or they can contribute to the selection of terms to be defined in a specialized dictionary, among other possible uses. There are no substantial differences regarding the definition of the notion of warrant. The authors seem to agree that these are theoretical-methodological criteria that guide the selection of terminology in all information contexts where subject representations are needed, and they are assigned as their main task to justify the inclusion, weighting or exclusion of terms.

The root of the issues faced in this work has a promising basis; in the last twenty-five years, the academic production on this topic has increased and has been greatly enriched, and in its diversity of approaches it has left a particularly fertile ground for the discussion of ideas. In the mentioned period, thirteen different warrants have been proposed, a regular work flow has been generated (both in the form of journal articles and conference papers) and the critical mass studying the warrants has significantly increased on an international level. However, a reorganization of the area is necessary. It is noted that there is a need to build an overall view, and to go in-depth in a series of issues that have been treated generically so far. It is also necessary to induce from what has been produced rules, premises, principles and methods that may be common to all warrants or that require their proper specification. Longer term works (monographs, books, postgraduate theses) that focus on this subject are also required.

On the way to a general conception of the warrants, this exploratory work has revealed—without exhausting the matter—different points of conflict, discussion or exchange of ideas around the warrants, and in particular, an important number of questions that could guide future research on the subject have been formulated. Perhaps the time has come when instead of thinking about expanding the list of warrants, it would be more productive to devote greater conceptual, methodological and applicative content to each of them, since they mostly have scarce literary warrant to support them.

The establishment of the relationship between warrants and epistemological theories is also relevant to promote a greater conceptual depth as well as to provide more support to field studies. Perhaps the most significant aspect to promote research in the area is that the topic of warrants remains particularly valid due to the importance assigned nowadays to subject retrieval on the internet, databases and databanks and other sources and information systems, linked to science, commerce, e-government, culture and entertainment industries. Their projection and utility are, therefore, sufficiently consolidated.

References


Barić, Mario et al. 2015. Diccionario de organización del conocimiento: Clasificación, indicación, terminología. 6th rev. ed. Montevideo: CSIC.
Appendix A: Typological table of warrants

<table>
<thead>
<tr>
<th>Name in English</th>
<th>Name in Spanish</th>
<th>Author and year</th>
<th>Comments</th>
<th>Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>Usage</td>
<td>Uso</td>
<td>Cutter, 1876</td>
<td>Antecedent of the user warrant</td>
<td>Empiricism</td>
</tr>
<tr>
<td>Literary warrant</td>
<td>Garantía literaria</td>
<td>Hulme, 1911</td>
<td>Antecedent of the academic warrant</td>
<td>Empiricism</td>
</tr>
<tr>
<td>Scientific/philosophical and educational warrant (consensus)</td>
<td>Garantía científico / filosófica y educacional-consenso</td>
<td>Bliss, 1929</td>
<td></td>
<td>Rationalism</td>
</tr>
<tr>
<td>Cultural warrant</td>
<td>Garantía cultural</td>
<td>Lee, 1976</td>
<td></td>
<td>Pragmatism</td>
</tr>
<tr>
<td>User warrant</td>
<td>Garantía de usuario</td>
<td>Lancaster, 1977</td>
<td></td>
<td>Empiricism</td>
</tr>
<tr>
<td>Logical warrant</td>
<td>Garantía lógica</td>
<td>Fraser, 1978</td>
<td></td>
<td>Rationalism</td>
</tr>
<tr>
<td>Request oriented warrant</td>
<td>Garantía orientada a la consulta o solicitud</td>
<td>Soergel, 1985, p. 230</td>
<td>Maybe a type of user warrant</td>
<td>Empiricism</td>
</tr>
<tr>
<td><strong>Name in English</strong></td>
<td><strong>Name in Spanish</strong></td>
<td><strong>Author and year</strong></td>
<td><strong>Comments</strong></td>
<td><strong>Approach</strong></td>
</tr>
<tr>
<td>--------------------------</td>
<td>---------------------------</td>
<td>---------------------------</td>
<td>-----------------------------------------------------------------------------</td>
<td>------------------------------</td>
</tr>
<tr>
<td>Semantic warrant</td>
<td>Garantía semántica</td>
<td>Beghtol, 1986</td>
<td>Generic name given to the literary, cultural, user and scientific/philosophical and educational warrants</td>
<td>Empiricism Rationalism Pragmatism</td>
</tr>
<tr>
<td>Gender warrant</td>
<td>Garantía de género</td>
<td>Olson and Ward, 1998</td>
<td>Maybe a type of cultural warrant</td>
<td>Pragmatism</td>
</tr>
<tr>
<td>Phenomenological warrant</td>
<td>Garantía fenomenológica</td>
<td>Ward, 2000</td>
<td></td>
<td>Empiricism</td>
</tr>
<tr>
<td>Structural warrant</td>
<td>Garantía estructural</td>
<td>Svenonius, 2000</td>
<td></td>
<td>Rationalism</td>
</tr>
<tr>
<td>Ethical warrant</td>
<td>Garantía ética</td>
<td>Beghtol, 2002b</td>
<td>Related to the cultural warrant</td>
<td>Pragmatism</td>
</tr>
<tr>
<td>Academic warrant (also named scholarly warrant)</td>
<td>Garantía académica</td>
<td>Sachs and Smiraglia, 2004</td>
<td>Similar to the scientific/philosophical and educational warrant</td>
<td>Rationalism</td>
</tr>
<tr>
<td>Organizational warrant</td>
<td>Garantía organizacional</td>
<td>National Information Standards Organization, 2005</td>
<td></td>
<td>Empiricism</td>
</tr>
<tr>
<td>Autopoietic warrant</td>
<td>Garantía autopoiética</td>
<td>Mai, 2011</td>
<td>Based on Rafferty and Hidderley, 2007. Maybe a type of user warrant</td>
<td>Empiricism</td>
</tr>
<tr>
<td>Textual warrant</td>
<td>Garantía textual</td>
<td>Tennis, Thornton and Filer, 2012</td>
<td></td>
<td>Empiricism</td>
</tr>
<tr>
<td>Market warrant</td>
<td>Garantía de mercado</td>
<td>Martínez Ávila, 2012</td>
<td></td>
<td>Empiricism</td>
</tr>
<tr>
<td>Indigenous warrant</td>
<td>Garantía indigenista</td>
<td>Doyle, 2013</td>
<td>A type of cultural warrant</td>
<td>Pragmatism</td>
</tr>
<tr>
<td>Genre warrant</td>
<td>Garantía de géneros</td>
<td>Andersen, 2015</td>
<td>A type of cultural warrant</td>
<td>Empiricism Pragmatism</td>
</tr>
<tr>
<td>Epistemic warrant</td>
<td>Garantía epistémica</td>
<td>Budd and Martínez Ávila, 2016</td>
<td></td>
<td>Rationalism Pragmatism</td>
</tr>
<tr>
<td>Policy warrant (corresponding to policy based indexing)</td>
<td>Garantía en políticas</td>
<td>Hjørland 2017,</td>
<td>A type of cultural warrant ?</td>
<td>Rationalism Pragmatism</td>
</tr>
</tbody>
</table>

Source: Barité (2018), revised, modified and expanded table for this paper.
Changing Perspectives on Classification as a Knowledge-Representation Process †

Barbara H. Kwaśnik

Syracuse University – School of Information Studies, Hinds Hall, Syracuse New York 13244-1100, United States, <bkwasnik@syr.edu>

Abstract: No matter how immutable a classification may seem, it is, after all, an artifact of the human imagination and functions in a particular place and time. The author describes her personal inquiry into classification as a knowledge-representation process. She traces her changing perspectives on how classifications should be viewed and evaluated by posing the following questions: 1) How does the classification process enable or constrain knowing about something or discovering something we did not already know?; 2) In what ways might we develop classifications that enhance our ability to discover meaningful information in the information stores that form a part of our scholarly as well as our everyday lives?; and 3) How might classifications mask or distort knowledge, and how might they serve to disenfranchise people and ideas? These questions are considered through a discussion of classification structures, personal classification, the link of classification to theory, everyday working classifications, translation of classifications, cognitive aspects, browsing, genres, warrant, and the difficulties of navigating complex ontological commitments. The through thread is the importance of context, because classifications can only be seen with respect to the human endeavors that generate them.

Keywords: classification, classifications, information, knowledge

† In writing about my research trajectory, I was repeatedly brought up by the awareness of how many people were involved. Here’s my opportunity to thank all of them: not only the groundbreakers who were, and continue to be, my intellectual inspiration, but also my professors and dissertation committee at Rutgers University, headed by James Anderson, as well as the colleagues on a day-to-day basis, the friends who bore my annoying habit of viewing just about any issue as a classification problem. I have been fortunate that my research and teaching at the School of Information Studies at Syracuse University have been closely intertwined, which is an increasingly rare privilege. The engagement of hundreds of students who brought so much to my classes was the source of my best examples and my deepest reflections. The International Society for Knowledge Organization and the Special Interest Group for Classification Research in ASIS&T were my home where I was welcomed as a young scholar and encouraged to develop my own trajectory. The influences and opportunities for a life of the mind were constant and enveloping. My most sincere thanks, though, go to my doctoral student co-authors, who participated with me in exploring interesting questions and helping me translate what we learned into the work summarized in this essay. My work on genres with Kevin Crowston was funded by an NSF grant: Crowston, Kevin (PI). How Can Document-Genre Metadata Improve Information Access for Large Digital Collections? NSF IIS Grant 04–14482, Jan. 2005 – Jan. 2007.

1.0 Introduction

We are incapable of “not” classifying. There are classifications everywhere, both formal and ad hoc; both enduring and ephemeral. After decades of studying and teaching I now understand better how classification both in and outside the library is not foremost about being tidy, but rather, about having a tool for seeing the world and understanding it. Our efforts at creating and using classifications serve to enable more efficient communication; they show patterns; they help visualize an overall view by showing clusters and areas of density and gaps. There are decisions made about “first cuts,” scope, definition, and relationships among the parts, all of this rendering a representation of some area of our lives and what we know about it.

My formal interest in classification goes back at least four decades, when as a beginner cataloger I realized that behind the library schemes used to organize resources physically on shelves was a world of intellectual richness and complexity that barely tipped the surface. The ques-
2.0 Knowledge and classification

I draw a distinction between merely observing, perceiving, or even describing things and truly knowing them. To know implies a process of integration of facts about objects and the context in which the objects and processes exist (Kwaśnik 1999, 23). It is not enough to know about the things but, rather, it is the relationship of one thing to another that creates the deeper understanding. We are familiar with how knowledge discovery and creation in the sciences may follow traditional processes, such as exploration, observation, description, analysis, and synthesis, as well as testing of phenomena and facts. It may also follow more interpretive paths, which are, nevertheless, based on evidence that is collected and appraised following consensual norms. All of this is conducted within the communication framework of a research community with its accepted methodology and set of techniques (Kwaśnik 1999, 23). Bronowski (1978) describes how even in the empirical sciences, though, the process is not entirely rational but is often sparked and then fueled by insight, hunches, and leaps of faith. Moreover, research is always conducted within a political and cultural reality (Olson 1998).

In my early thinking in the 1980s, I was drawn to the classification work in the sciences, not because I thought it was worthier of being prioritized, but because it is linked to underlying theories or conceptual frameworks. This link to theory seemed a very important one to me since it rendered the classification as somehow more substantive, more enduring. I have since modified this view, but at the beginning of my career I assumed that the more theoretical a classification was, the better it was at representing knowledge. I agreed with Kaplan (1963) who said that theories and models are a “symbolic dimension of experience as opposed to the apprehension of brute fact” (294). In the same way, I thought a classification that is thrown together without any conceptual glue to hold it together is typically wobbly as a knowledge-representation structure. Overall, I believed that, indeed, a classification itself could be construed as a theory—a structure of entities and their specified relationships (Kwaśnik 1994). Whether it was good or flawed depended on the value of the conceptual scaffolding.

Following on this analogy of classification to theory, I observed that the process of classification can be used in a formative way during the preliminary stages of inquiry as a heuristic tool in discovery, analysis, and in theorizing (Davies 1989). Once concepts gel and the relationships among them become understood, a classification can be used as a rich representation and is thus useful to communication and in generating a fresh cycle of exploration, comparison, and theorizing. Kaplan (1967) states that “theory is not the aggregate of the new laws but their connectedness, as a bridge consists of girders only in that the girders are joined together in a particular way” (297). I believed that a classification works in much the same way, connecting concepts in a useful structure. If successful, it is like a theory, descriptive, explanatory, heuristic, fruitful, and perhaps also elegant, parsimonious, and robust (Kwaśnik 1994).

This led me to believe that classification is somehow a fundamental aspect of nature, sort of like the Fibonacci Sequence, and that the “lawful” integration of theory and classification yielded the most robust schemes. There are examples to suggest this, the periodic table of the elements being one. Here we have a classification that has endured through different theoretical explanations and continues to be useful to this day. The scheme, which started in the nineteenth century as an observation of the regular change in atomic weight among elements, eventually yielded a pattern that led to the discovery of new elements, and through the lens of increasing theoretic understanding an explanation of why this occurred. From my perspective, it was a comforting thought that while the principles of a classification could evolve, the underlying concept remained solid.

I have come to believe that this might be a dangerous belief in the sense that not everything sorts itself out as beautifully as the elements in the periodic table seem to do. In fact, deeper analysis reveals beyond a doubt how any classification, no matter how immutable it may seem, is after all an artifact of the human imagination and functions in a particular place and time. For many phenomena there are many ways to interpret, visualize and explain them, even my revered periodic table. The value of any one approach depends on many factors, including the context in which the classification is being invoked. Putting up any one classification as ideal implies that classification schemes that do not measure up are somehow inferior rather than simply different, or that they do not have a useful purpose. While many traditional classifications have strength and merit, I have
come to realize the exclusive embrace of elite classifications and classificatory structures may build in subtle assumptions and biases, making it more difficult to admit diverse perspectives.

3.0. Classification structures

When I first investigated classification as a research topic the question I posed was, “How do we systematically evaluate a classification?” I wanted to step back and impartially analyze what made a classification tick—any classification. What are the features that provide strength in terms of knowledge representation, and what features misguide or interfere with it? I was additionally motivated by my need to develop techniques for teaching about classification in a way that gave students a toolkit they could use to describe and evaluate any classification that came their way in a thoughtful and careful way, especially the legacy tools they would be using in their work. An important part of this profile was a description of the structural properties of the classification, how the parts worked together in terms of relationships. There are many kinds of classificatory structures, but for this review I will discuss the two that are probably most familiar: hierarchies and trees.

3.1 Hierarchies

I start with hierarchies, because they are perhaps the most recognizable (and perhaps the most misunderstood). People call all kinds of things hierarchies, but here I refer to the logical structure we have inherited from Aristotle (1963). His view was that, after careful observation, one could learn how things could be aggregated and differentiated “naturally.” It assumed the division and aggregation of classes would be valid, because one had to arrive at the “essential qualities” of what was being classified. Yes, we now question that such structures require us to choose one ideal representation over possible others, but nevertheless, hierarchies are often sought out as the structure of choice for their many strengths described below.

Essentially, a hierarchy is a structure with a top class that defines the scope of the classification. The top class includes all its subclasses and sub-subclasses. A true hierarchy has only one type of relationship, the “species/differentia” relationship, also known as the "generic" relationship, or informally the “is-a” relationship. The strict requirement for inclusion ensures that what is true for the top class is true for all the subclasses. This property is called “inheritance,” that is, attributes are inherited by a subclass from its superclass. “Transitivity” is an important outcome of this carefully controlled structure, because one can assume that all classes are members of not only their immediate superclass but of every superclass above that one.

Hierarchies have predetermined, predictable, and systematic rules for association and distinction. There are “necessary” and “sufficient” conditions for when something may belong to one class and be distinguished from another class. The rules attempt to use the most “essential” type of information for distinguishing one class from another, and thus, entities differ from sibling entities in a predictable way. Often the rules are based on some theory that is the foundation for the classification. Finally, a hierarchy invokes the rule of “mutual exclusivity,” which means that each entity can belong to one and only one class.

To take one example, in western medicine, concepts lend themselves to hierarchical arrangement. Here’s an abbreviated snippet from the National Library of Medicine’s Medical Subject Headings (MeSH) (https://meshb.nlm.nih.gov/record/ui?ui=D013494).

Nervous System Diseases [C10]
  Central Nervous System Diseases [C10.228]
  Brain Diseases [C10.228.140]
  Basal Ganglia Diseases [C10.228.140.079]
    Basal Ganglia Cerebrovascular Disease [C10.228.140.079.127]
    Chorea Gravidarum [C10.228.140.079.294]
  ...
  Parkinsonian Disorders [C10.228.140.079.862]
  Supranuclear Palsy, Progressive [C10.228.140.079.882]
  Tourette Syndrome [C10.228.140.079.898]

In this example, searching for the rare disease progressive supranuclear palsy, we learn that it is part of several linked hierarchies. One is shown here, under nervous diseases. The information flows in many directions: from the top terms down, from the “sister” terms, and also from the related terms (such as movement disorders) in other parts of the schedules. It helps the searcher identify the landscape. Knowing that PSP is located near parkinsonian disorders and tourette syndrome helps define the nature of the disease, and indeed, helps explain why it is difficult to diagnose in its early stages.

It is obvious why such careful constructions are appealing in knowledge representation. If valid in their underlying assumptions and definitions they offer complete and comprehensive information. The affordance of inheritance provides for an economy of notation, and perhaps most important, a hierarchy offers the ability to make inferences from incomplete information. For instance, one can infer that a female kitten is, like all cats, a mammal, and by her essential mammalian features could be expected to eventually bear live young and breast feed them, even if she is not at present doing so. Put another way, the hierarchy succinctly carries a great deal of information that can
be used to represent the domain, to explore, and to provide conceptual fodder for further discovery. Above all, a hierarchy is built on logical principles, so to many people it seems trustworthy. The parts must fit together; it should be comprehensive and not have loose ends. Building and maintaining hierarchies requires a strong *à priori* conceptual framework as well as consensus to guide the development of the rules. That is why hierarchies are often deductively created, rather than built from the bottom up.

Not all knowledge domains lend themselves to a strictly controlled hierarchical representation, though. Hierarchies are problematic for a number of reasons. First and foremost, we acknowledge that it is often very difficult to identify the ideal “essential” partition points in any domain of knowledge. Many phenomena can be seen from several perspectives, depending on the context and the goal of the classification. There may be multiple and diverse criteria, and there may be a question of which criterion to invoke first. For example, the traditional classification of animals divides them up into ever more specific taxonomic ranks, from kingdom to species. This order precludes the consideration of the differences between animals in the wild and animals in captivity. Moreover, it is awkward to use this classification to represent ecological systems gracefully, since there are many other factors to consider besides the animal’s morphological “essence.”

Finally, a true hierarchy requires deep knowledge and consensus about the domain in order to determine the rules for defining classes, partitioning, and aggregation. If there is no conceptual framework guiding these choices, the classification can seem incoherent or contradictory. Thus, in new and emerging fields where knowledge is incomplete, it is sometimes unwise to commit to a hierarchical classification. Even when the field is mature, though, but rapidly changing due to new incoming information, maintaining a hierarchy can be dicey. For instance, we are familiar with the muddy classification of heavenly bodies, such as planets.

The principle of transitivity and inheritance requires that all the entities in a hierarchy be at the same level of conceptual granularity, thus a hierarchy does not accommodate differences of scale for the same phenomenon. For example, a beach might be construed as a kind of land mass as you might see it from space, or it might be a kind of habitat, or it might be viewed as an aggregate of materials, such as sand. A hierarchy encompassing all the views in one structure would be impractical and confusing in terms of making inferences and comparisons.

### 3.2 Trees

Trees are another familiar type of classificatory structure. A tree divides and subdivides its classes just as in a hierarchy, but the relationship among the classes is not necessarily generic. There are many types of relationships possible, such as part/whole, a kinship tree, or an organizational chart. In a part/whole scheme such as:

- North America → United States → NY State → Onondaga County → City of Syracuse

you can see that Syracuse is a part of Onondaga County, which is part of NY State, and so on, but Syracuse is not a “kind of” county, nor is the county a “kind of” state. This structure limits the transitivity of information, because what is true of North America does not carry down to the city level in terms of shared “essential” features. Similarly, if an engine of a car contains pistons, spark plugs, and valves, you do not have a great deal of information about the relationship between pistons and spark plugs except to know they are both part of a car engine. They can, in fact, be totally different entities. What unifies them in the scheme is their position in the engine.

In a traditional kinship tree, you may describe the flow of who begat whom, but a daughter is not a kind of mother, nor is a mother a kind of grandfather. Instead, the relationships are determined by blood and legal affiliation. In an organizational chart there is a clear purpose, and that is to show “who reports to whom” or perhaps “who is managed by whom.” This is not to say that kinship trees or organizational charts do not yield a great deal of information, but they are not as rich and inclusive as the classic hierarchical structure in terms of showing the unity of the whole. Instead, one or two critical relationships are highlighted, which makes them easy to comprehend and analyze along those relationships.

In summary, both hierarchies and trees are useful systematic knowledge-representation structures with different properties and strengths and different constraints. In both you must know about the domain to pre-determine the first cut and the rules for membership in classes. Both kinds of classification structures have many challenges in being applied, though, because the realities of application do not always map well onto the requirements of such structures.

### 3.3 Facetted classification

My growing awareness that classificatory thinking was both culturally and psychologically influenced led me to further explorations. Among these was facetted classification, which is not a different representation structure, but rather an approach that allows the classifier to view the world as dynamic, and indeed, provisional, in how it is constructed. The approach is credited by many to S.R. Ranganathan, an Indian scholar, who posited that any entity could be viewed from a number of fundamental perspec-
tives or facets. He suggested that these are: personality, matter, energy, space, and time (Ranganathan, 1967). While the discussion of the nature of these facets varies in his work, the principle has caught on and endured. The notion that any entity can be analyzed into aspects, each representing some feature or quality, freed up designers to create schemes that were multidimensional. Note this is not the same as breaking down an entity into components, but rather viewing the same entity from different perspectives—same object different views. One of the clearest examples that explicitly built on Ranganathan's principles is the *Art and Architecture Thesaurus*, a compilation of vocabulary for the indexing and retrieval of objects and literature on material culture, which in its diversity defies easy description and classification into any one classificatory structure. The *A&AT* allows the creation of a string to represent a topic or object using the core categories of period/style, place, process, material, and object. For example:

19th Century Japanese raku ceramic vase
*Arts & Crafts American oak desk*

In doing so it is then possible to search by any one of the components (e.g., all things “arts & crafts,” or all “vases”), or in combination. It also allows for the graceful addition of new objects and topics, so long as they can be analyzed using the five categories.

This approach has extended well beyond formal collections and is very popular in shopping sites as well as visualized analyses of all kinds. Not all use Ranganathan’s principles, but the result is essentially the same. Each facet can be developed following its own logic and structure, and then synthesized into expressive strings. Obviously, the advantages are you do not need exhaustive knowledge so long as you can identify important fundamental “aspects.” It is a hospitable and flexible approach without the need to have a strong, immutable theory for the scheme overall. At the same time, it can accommodate a variety of theoretical structures and models in the facet components, and, most important, it can sustain a variety of perspectives. Thus, a flower can be considered as food, as a feature of specific habitats, as a commodity, and so on.

While faceted schemes have pragmatic appeal, there are some things to consider. First, is the difficulty in identifying the core facets. They should be robust and so, while complete knowledge of the domain is not necessary, enough must be understood to accommodate all new entities. For example, you might view the traditional classification of instruments as a faceted scheme, incorporating material, process of creating sound, origin, and so on, but the scheme hits a speed bump when you want to also include electronic instruments. Second, while it is freeing not to have a required binding conceptual framework, this also means that the scheme remains essentially descriptive. There is no guidance for how to read or interpret the relationship “among” the facets. Finally, it may be difficult to visualize all in one grand picture. A scheme might include a timeline, a hierarchy and a tree, but no built-in clue on how these should be presented. Even so, chances are if you look around at modern classification, it will likely as not be a faceted scheme (Kwaśniki 1999, 39-43). In much of my research I’ve employed the notion of faceted classification, especially in the challenges presented by multidimensional situations.

### 4.0 Practical classifications

The study of formal classification led me to an appreciation of their formidable power, but also piqued my interest in simple, practical classifications that perhaps deviated from the “ideal.” Many of these exist to help with tasks such as shopping, diagnosis, or description without being necessarily bound by a particular theoretical framework. That is, they are not without an underlying rationale, but they do not purport to be “true” or enduring. They are simply there to organize some phenomenon in a useful way. After having scoffed at such schemes I became fond of them, because they demonstrate our human ability to be in touch with the power of a good classification that uses visualizations and simple metrics to help navigate through more complex information. The point of the following examples is that in many ways they provide for enhancement of description and searching where sometimes the formal classifications are lacking or overly complicated. These classifications demonstrate that iterative, flexible design, combined with other search features can be very effective and certainly easier to maintain.

#### 4.1 Keys

My trusty old Peterson’s field guide to wildflowers (1974) is an example of such a pragmatic classification that makes identification of wild plants more accessible. It is a key, which is a type of classification that chooses one or two obvious features to lead into the more formal classification. In this case, the plants are organized by petal color, a feature not “essential” in the Aristotelean sense but recognizable. Then, by icons and clear descriptions, the identification can proceed further. In other words, the key is just that, a key. The design of such a key might be frustrating to the botanist because it might seem superficial, but to a novice it is a testament to the communicative power of such tools.
4.2 eBay.com

Commercial websites use classifications in contexts where the content may be in constant flux, the user population is unknown, or if it is known, we can assume it is diverse; and where it is desirable that the classification be very simple and straightforward so that all levels of users can learn it. For instance, eBay.com maintains a classification of millions of objects. It fails miserably when analyzed using the formal criteria of coherence, but it is a classification of current objects—everything on eBay exists and is for sale—it is not meant to endure forever. Despite its rather sloppy design, it is surprisingly robust and hospitable. When you consider that it reflects the terminology used by several million people for an amazingly wide and constantly shifting array of items, it is quite impressive. In terms of accessibility, there are very few terms in the main categories that are difficult to understand. One of the strong points is the meshing of the classification with many other access strategies. If the classification fails, there are other avenues to pursue (Kwaśnik and Liu 2000).

4.3 Amazon.com

Along the same lines, when I studied the amazon.com book division years ago (Kwaśnik, 2002), even back then the affordances of multiple access points ensured that “something” would be found no matter what the user entered. The classification achieved a multi-perspective view allowing for a facetted approach, and if one approach did not work another was readily available. I concluded (284) that “in general, amazon.com’s scheme can be viewed as more pragmatic and enumerative than as based on a model of knowledge.” It is a classification meant to encourage buying and uses as many routes to the goal as possible, including a simple but redundant vocabulary without much attention to structural integrity but able to provide a rich network of subjects.

4.4 Scientific and naïve classifications

For a final example, we explored the idea of “teaming up” scientific and naïve classifications. We compared two separate but related classification schemes in the area of medical information to better understand how they might be used together and inform one another. We contrasted MeSH with the consumer health website, WebMD.com. Using the term “autism” we compared the strengths and limitations from the perspective of vocabulary, syntax and classificatory structure, context, and warrant. We conclude that in terms of vocabulary and concepts, MeSH may benefit from WebMD’s approach to ongoing updates and currency as well as the contextualization of terms. At the same time, WebMD.com may benefit from some form of vocabulary control for richer expansion of terms and archival retrieval (Kwaśnik and Flaherty 2010).

5.0 Challenging the Aristotelian paradigm

As mentioned, two foundational but possibly conflicting assumptions took shape in my early pursuit of studying classification: the first was my firm belief that formal classificatory structures, such as hierarchies and trees, help advance understanding because of their ability to represent not only the elements of a domain but also the relationships among them, thus yielding knowledge structures that not only describe but also explain. At the same time, I realized that people in their enactment of classification brought their own contextual understanding to them and the two did not always map well.

5.1 Accounting for context in the classification of personal documents

In my dissertation, The Influence of Context on Classificatory Behavior (Kwaśnik 1989b, 1991), I wanted to see how people create and then utilize classification, that is, I wanted to learn about personal information management in everyday life. I interviewed university professors in their offices and recorded their documents and the organization of these documents on shelves, in drawers, on the computer, in their briefcases, taped to the door, and in various piles and files, always in their own words. The findings, in a nutshell, were they did not organize things as they are organized in formal library collections by subject and form. In fact, the contextual factors played a critical role—factors such as the purpose of the document or its currency. Documents with the same content or subject could be classified differently depending on how they would be used (Kwaśnik 1989a). The findings suggested that while formal classifications exist and prove very useful, the establishment of universal schemes is more problematic in situations in which context plays a part. So, formal logical representation works when the domain is well understood and there is consensus on the underlying conceptual structure, but what about all the rest?

5.2. Influences of cognitive anthropology

Along the way to finding a conceptual framework for my dissertation I was introduced to the field of cognitive anthropology, and most intriguing to me, category choosing. In a hierarchy, each member of a class is an equally good representative of that class, since each member must possess all the requisite attributes. In a pure hierarchy an entity cannot be “sort of” in a class. The boundaries and rules for inclusion and exclusion are defined and predetermined.
An entity cannot belong to more than one class following the principle of mutual exclusivity. These traditional assumptions came increasingly into question, because anthropological and cognitive evidence from studies of humans did not support them. The formal properties do not necessarily map accurately to our human cognitive processes, that is, we do not all store our concepts in hierarchies. We know that humans have fuzzier notions of what constitutes a boundary on a class of things. These boundaries may change with circumstances and the experiences of the classifier.

5.2.1 Prototype theory and the principle of family resemblances

I was influenced in particular by researchers such as Eleanor Rosch whose prototype theory posits that not all members of a class may be perceived as equally good representatives of that class. Not all birds fall naturally into the class of “birds;” some seem more birdlike and others less so. She posited that for every class, some objects become prototypical and form a best example of that class (Rosch 1973). Another notion that influenced me was the idea of family resemblances (Rosch and Mervis 1975), where they argue that other notion that influenced me was the idea of family resemblances (Rosch and Mervis 1975), where they argue that the principle of family resemblance can be construed as “a logical alternative to criterial attributes.” They were arguing against (603) “a tenacious tradition of thought in philosophy and psychology which assumes that items can bear a categorical relationship to each other only by means of the possession of common criterial attributes.” Their study presented empirical confirmation that formal criteria are neither a logical nor psychological necessity. This means that for any given class, certain attributes define members of that class, but not all members must possess all the attributes nor exhibit them as strongly (thus defying the principle of necessary conditions). Imagine in your family there are recognizable family traits, but they are not distributed equally among everyone. The idea of family resemblances raised some interesting questions, such as which of these attributes is defining? Are all combinations defining?

5.2.2 George Kelly and personal construct theory

Another important contribution to understanding the cognitive aspects of classification was George Kelly’s Personal Construct Theory (1955, 1970). Kelly posited that everyone construes the world in a different and individual way. His original work included an intriguing appendix: The Repertory Grid. This tool eventually was used outside its original intention and became popular for making people’s individual implicit constructs explicit. Building on the findings from my dissertation that variations in naming were large, we used the Repgrid technique to explore the naming of office documents. The study yielded a fine-tuned descriptive analysis of consensus, conflict, and correspondence among people for common documents, demonstrating that in fact perfect correspondence in naming between individuals is not the norm (Kwaśnik and Jörgensen 1992).

6.0 Extending the borders

In the time I have been studying classification, we have seen a shift to unification and standardization of bibliographic systems, not just in the United States but also globally. This means that traditional classifications, originally designed in a particular country or for a particular collection are now being stretched, in Michèle Hudon’s words (1997), to cover cultural and linguistic artifacts and concepts quite different from those originally intended. This had special significance for me, because given my understanding of a classification’s structure and impact, I knew that extending them was not simply a matter of one-to-one translation.

6.1 Translating classifications

In a study comparing the Dewey Decimal Classification and the Korean Decimal Classification, two bibliographic schemes from different cultures, we found that obvious differences could be accommodated (Kwaśnik and Chun 2004). For instance, the DDC emphasized Christianity, while the KDC allowed more room for Buddhism. The KDC offered greater expressiveness for terms such as “calligraphy.” The differences that were more profound, however, were those that construed subjects very differently. For example, “war” is treated as a social process in the DDC, and is placed near diplomacy, whereas in the KDC it is classified as a social problem and is near suicide (197). Such a difference in conceptual mapping makes culturally sensitive translation challenging.

Translating the vocabulary of a classification has the typical issues of translation in general. In a study of people’s use of even the very most basic kinship terms such as “mother,” there are many problems. Among these are finding corresponding terminology and being able to reflect the relationship between terms in the target language correctly. It is surprising how many denotations and connotations the term “mother” has, even in English. We found in the process of translation there may be structural shifts; some terms may have broader definitions and some narrower. There may be differences in how similar terms are construed, and there may be additional criteria of distinction (such as birth order in the case of kinship systems). We suggested that not only terms themselves but also inter-term relationships need to be preserved in cross-cul-
tural, cross-lingual classification translations so that both the source and the target schemes are truly reflective (Kwaśnik and Rubin 2003).

7.0 The importance of context

It was evident to me that a key ingredient in making classifications more responsive and resonant was to find some way of incorporating context into the process. A professor organizes office documents with an eye to the potential uses. A person browsing a collection brings to it personal insights or needs and uses those to help navigate the space. The situation in which classificatory decisions are made plays an important part. Yet, it is quite difficult to reconcile rigid classification schemes with infinitely individual ones. Starting with the findings of my dissertation, the dilemma of creating and using classifications that are accommodating of many perspectives always seems to boil down to one important factor: context. Context defines the scope and the vocabulary. It decides on the elements themselves and which classificatory relationships are pertinent. The following examples show two streams of research in which I tried to find ways of identifying and then representing contextual factors.

7.1 Context and discovery: browsing

One of the features of a classification, any classification, is that it creates affordances for exploration. A classified environment can be searched or browsed for something even if we only suspect or expect it will be there but do not know for sure. Browsing is a method of information seeking that allows the user to explore and navigate without having to specify a query. As such, it is a good way of dealing with an unfamiliar environment or with multiple options or choices. In this way, browsable systems can be invaluable to users crossing over into a new and unfamiliar domain. Browsing reduces cognitive load, because it is generally easier to recognize something once it is viewable rather than to recall a term for it. As well, a key feature of browsing is the ability to hold several parallel paths at once without having to commit to just one.

I wanted to investigate what people do when they browse. The term had been variously defined as searching, scanning, navigating, skimming, sampling, and exploring. It was often described as searching “without a particular purpose” and without a set structure as compared to a database search, for instance. We conducted some informal observations of people browsing in catalogs, online, at a farmer’s market, and so on (Kwaśnik and Yoon 1990). The purpose was not simply to record what they did or what “nodes” they visited and how often, but more fundamentally to identify what function they accomplished. Ultimately, we hoped this would provide a set of principles for designing browsable interfaces.

The studies showed that with respect to the structure of the environment, the notion of an unstructured environment is probably not as useful as observing what structures are perceived and how they affect behavior. People will create structures even amidst seeming chaos. Similarly, they will develop a purpose to the process, even if they seemingly started off without one. Comparisons and strategy are developed iteratively. This amazing human capability can be described by the following functions: orientation, place marking, identification, resolution of anomalies, comparisons, and transitions. Each of these functions is performed by constant interaction with the browsing environment, but also with past experiences, future plans, and many other factors the browser brings to the experience. Thus, we can say that browsing is not a passive activity, because there is a formidable amount of sense making involved. As a way of coping with the browsing environment, the browser is constantly devising classifications or views based on a shifting context. Being able to harness these abilities would make interfaces easier and more productive (Kwaśnik 1993).

7.2 Genres

My growing awareness that classification of any kind is a social act led me to the study of genres as they play out in knowledge representation for information seeking and use. A genre identifies something as an integrated cluster of features enacted in a social environment. My colleague, Kevin Crowston, and I conducted a series of studies to see if identifying the genre of a document would improve information access in large digital collections through the identification of document genre as a facet of document and query representation. For instance, knowing something was a computer program might help distinguish it from a musical program, each of these being a different genre. Because most genres are characterized by both form and purpose, identifying the genre provides information as to the document’s purpose and its fit to a user’s situation, which can be otherwise difficult to assess (Crowston and Kwaśnik 2003, 2005).

First, we needed to define genres for ourselves since this is a very old area of study and crosses many disciplinary lines, from the arts to business. Genres are a way people refer to communicative acts that is understood by them, more or less, but is often difficult to describe in its particulars. Thus, genres are recognized and used, but not so readily described and defined. In our work, we drew on the definition of genre proposed by Yates and Orlikowski (1992, 543), who describe genre as “a distinctive type of communicative action, characterized by a socially recognized communicative purpose and common aspects of form.” Note this does not
mean that a genre can be seen purely as a set of document attributes, making the representation of genres a complex and difficult proposition.

Among other things, we wanted to know how people talk about the genre of documents. How do people make use of new, unnamed, and emerging genres? What clues do people use to identify genre when engaged in information-access activities? What facets (basic attributes) of genres do people perceive (Crowston and Kwaśniik 2003)?

Our plan was to create a taxonomy of genres by studies of people searching for information in the field. This taxonomy would be used to create a simulated search situation in which we could observe the difference between searching with the aid of genre information and without.

Our assumption going in was that a faceted scheme for genres would be best given their multidimensionality and complexity (Kwaśniik and Crowston 2004). We attempted to harvest clues people used to identify genres, such as “scholarly language” or “reverse chronological dated content” and then reduce them by analysis into possible genre facets. The clues and resulting user-generated scheme were not possible, because the concept of genre was even more slippery than we anticipated. We had difficulty defining genres and developing the scope and expressiveness of the scheme from what our participants told us. They were not able to reliably identify the genre unit or provide unambiguous genre labels. When prompted, they found it difficult to identify genre attributes. There were challenges in distinguishing form and content, as well as challenges in identifying purpose. Finally, the granularity of their tasks differed immensely creating imbalances in the granularity of the terms we could use (Kwaśniik, et al. 2006). As a result, we worked around the lack of a user-generated faceted view of genres and created a researcher-compiled working taxonomy for the purposes of the experiments (Crowston, et al. 2011). In the end, we were not able to demonstrate a substantive change between plain searches and those enriched with genre information, but I believe the full potential of genre representation remains to be explored.

8.0 Classification at the intersection with human endeavor

*Sorting Things Out* by Geoffrey C. Bowker and Susan Leigh Star (1999), made an enormous impression on me and on the knowledge organization community and beyond. In this work, the authors examined revelatory classifications and standards to show how such classifications silently influence the infrastructure of information, affecting not only policy, but also our daily lives. They were not the first to urge that we question classifications, but their insights were profound, vivid, and compelling. They showed how the system of apartheid, for instance, embodied the pain of South Africa’s history, or the classification of tuberculosis affected people’s life trajectories.

Classification schemes reflect the knowledge of the domain being classified but also the perspective of the classifier, thus no classification can ever by understood out of context. While we take for granted that classifications do have a social impact, it is not always easy to say precisely how, although we can certainly feel the effects. Potential marginalization, rules for inclusion or exclusion, labeling and naming are all outcomes of classification decisions. Those in power design the classification and then have power over those who are not able to change it. The news is full of examples on a daily basis, from pressure on the Library of Congress to change the term “illegal aliens” to “undocumented immigrants” to who can use the term “champagne.” Political resistance often means changing the ruling classification. Many standards are based on classification. Many conflicts have at their core a dispute over basic classifications: when does life begin and when does death occur? What makes a crime a crime? What defines a country? In learning how to evaluate a classification we should always take the critical view. Who devised it? Whose purpose is being served?

8.1 The case of ontological commitment and warrant

*Sorting Things Out* engendered a critical eye with respect to my analysis and perception of classification, but it was one thing to find the strengths and flaws and another to develop a vocabulary for discussing this systematically and coherently. Fortunately, I was asked to contribute to a festschrift for Claire Beghtol (Kwaśniik 2010) and chose to focus on her pivotal and far-reaching 1986 article “Semantic Validity: Concepts and Warrant in Bibliographic Classification Systems” (Beghtol 1986). In this article she explores the semantic, rather than the syntactic axis of bibliographic classification systems. According to her, the attention of scholars on faceted schemes and classificatory structures had heretofore pulled our attention to the syntactic aspects (e.g., concept division and citation order), with semantics being considered more or less a question of the terms and their relationships and somewhat taken for granted. In this paper she states (110-11) that “the warrant of a classification system can be thought of as the authority a classificationist invokes first to justify and subsequently to verify decisions about what class/concepts should appear in the schedules … The semantic warrant of a system thus provides the principal authorization for supposing that some class or concept or notational device will be helpful and meaningful to classifiers and ultimately to the users of documents.” Warrant emerges from various points of authority: literary warrant, scientific/philosophical warrant, educational warrant, and cultural warrant, each with its own effect in terms of establish-
ing the semantics and then also the syntax of any given classification (119-221).

This was a revolutionary idea in the sense that notions of meaning being fixed have guided the design of many of our systems, because it was assumed that meaning became more stable and consensus firmer as the evidence mounted and the ideas withstood the test of time. Yet, modern approaches assume that meaning is not fixed and is created in use. It is also interesting to consider contemporary phenomena such as wikipedia.org, where the classification and the content are built cooperatively. That is, in principle, both the text and the classification that organizes the texts in such emergent systems are not managed from the top. Nobody questions the fact that such systems must be flexible and dynamic, and yet nobody wants an amorphous mess either. Our challenge is to assess the warrant for any given classification project and judge the classification against it.

One example of such a challenge is the classification of academic departments and programs. In the modern American university, there is often a federated system of schools and colleges, each with its own warrant for how it describes and labels the knowledge in its purview, what Elaine Svenonius called ontological commitments (Svenonius 1997). Chemistry views its own world differently than physics. Each of these comes with its own ontological commitments and its own body of knowledge. The forensics student’s program of study is not based on the supporting and contributing disciplines, however, but rather on a prescribed sequence of professional practice: identification of crime; collection of evidence (autopsy, traces); analysis of evidence; and support of the preparation of a legal case.

My takeaway from studying these cases is that, broadly speaking, when classification is structured to support human endeavors, the purpose is different than when it is structured to support science. Thus, understanding the underlying warrant is all the more important.

9.0 Full round back to hierarchies

In a way, then, my early respect for hierarchies would seem to have been validated up to a point, if, and only if, the circumstances supporting a hierarchical structure were evident. Recently, though, even this qualified view has been somewhat shaken. Hope Olson’s mission is to analyze our traditional knowledge representation systems from the point of view of those whose voices are not well reflected. In her article “How We Construct Subjects” (2007), she takes apart the notions behind hierarchies and brings to bear feminist thinking to offer a penetrating critique. She posits that hierarchies are by nature flawed because they require one element to be in the superior position and all other elements subordinate to that. This structure creates skewed assumptions that privilege one set of elements over others. I will use my own example here: imagine a hierarchical classification of astronauts. At the top is the term “astronauts.” On the next level down are subclasses of astronauts: “minority astronauts,” “women astronauts,” and so on. This may seem like a laudable effort at inclusiveness. There is no subclass for “men astronauts,” however, because the notion of astronauts being men is the default and is baked into the assumptions. In the chain of transitivity, men hold the defining set of attributes. This is a dilemma, because while one would like a way to represent the special attributes of women astronauts, placing them in the subordinate position means that they are defined by the male criteria first and foremost. From my perspective, there does not seem to be a good way to undo this imbalance in a hierarchy.

Having laid out the limitations both in content and structure, Olson suggests rewriting and restructing our schemes so that the all-important connections are visible—a web instead of a hierarchy (522). According to her (522), we need “richer and more situated logical models” that allow for the representation of interdependence and connectedness. I am just now beginning to rethink how my favorite classifications could be reframed in this way, or if they even should be. The power of hierarchies and other formal classifications is not easily dismissed, but at the same time the fact that they are so embedded in our culture should be explored. What we see as taken for granted could be hiding subtle and not so subtle biases.
9.0 Summary

Classification is beautifully recursive. What we know guides our classifications, and in turn, our classifications guide what we are able to know. Many questions remain: 1) Who creates the classifications by which we must all live?; 2) Who has the authority to change them?; and, 3) What is an effective way of creating classifications that are inclusive but also effective? We use classifications to better capture what we know; we also use them to embody our values and perspectives. We don’t have a choice of whether to classify or not, but we are obliged to pay attention to the consequences of what we do.

References


Knowledge Organization

Sixth Annual “Best Paper in KO Award” for Volume 45 (2018)*

DOI:10.5771/0943-7444-2019-8-668

Awarded to:


Abstract: Traditional knowledge organization approaches struggle to make large user-generated collections navigable, especially when these collections are quickly growing, in which currency is of particular concern, for which professional classification design is too costly. Many of these collections use folksonomies for labelling and organization as a low-cost but flawed knowledge organization approach. While several computational approaches offer ways to ameliorate the worst flaws of folksonomies, some user-generated collections have implemented a human judgment-centered alternative to produce structured folksonomies. An analysis of three such implementations reveals design differences within the space. This approach, termed “curated folksonomy,” presents a new object of study for knowledge organization and represents one answer to the tension between scalability and the value of human judgment.

Julia Bullard is an Assistant Professor at the School of Information at the University of British Columbia. She holds a PhD in information studies from the University of Texas at Austin, an MLIS from the University of British Columbia, an MA in cultural studies & critical theory from McMaster University, and a BA in English literature from St. Francis Xavier University. She teaches courses in information organization, metadata, and social justice. Her current research projects focus on how terminology systems depict LGBTQ2IA+ subjects and express theories of national values.

Books Recently Published
Compiled by J. Bradford Young

DOI:10.5771/0943-7444-2019-8-669
