Contents

Editorial
Richard P. Smiraglia.
ISKO 16’s Bookshelf: Knowledge Organization on the Verge of the Pandemic—An Editorial .......................... 619

Articles
Deborah Lee.
Reality, Unreality, and Artistic Deception:
The Ethical Dimensions of the Knowledge Organization of Art Documentation ................................................... 631

Isidoro Gil-Leiva, Mariângela Spotti Lopes Fujita, Pedro Díaz Ortuño and Daniela Majorie dos Reis.
The Abandonment of the Assignment of Subject Headings and Classification Codes in University Libraries Due to the Massive Emergence of Electronic Books .......................................................................................... 646

Brian Dobreski, Jian Qin and Melissa Resnick.
Depicting Historical Persons and Identities:
A Faceted Approach .................................................................................................................................................. 668

Maria Luísa de Almeida Campos, Linair Maria Campos and Nilson Theobald Barbosa.
The Challenges of Semantic Interoperability in the Era of eScience on the Web ....................................................... 680

Reviews of Concepts in Knowledge Organization
Matthew S. Mayernik.
Metadata ........................................................................................................................................................................ 696

Seventh Annual “Best Paper in KO Award”
for Volume 46 (2019) ........................................................................................................................................... 714

Books Recently Published .................................................................................................................................. 716
ISKO 16’s Bookshelf: Knowledge Organization on the Verge of the Pandemic—An Editorial

Richard P. Smiraglia

Institute for Knowledge Organization and Structure, Inc., Lake Oswego OR 97035 USA
E-mail: richard.smiraglia@knoworg.org

Richard P. Smiraglia is Senior Fellow of the Institute for Knowledge Organization and Structure in Lake Oswego, Oregon. He is a Guest Fellow of the Data Archives and Networked Services of the Royal Netherlands Academy of the Arts and Sciences, Professor Emeritus of the University of Wisconsin-Milwaukee and Long Island University, and retiring (with the present issue) Editor-in-Chief of the journal Knowledge Organization. His work explores ontology extraction and evolution of knowledge in domains, classification interaction, classification-based knowledge maps, the cultural role of authorship, the representation of knowledge in knowledge organization systems and the phenomenon of instantiation among information objects.


Abstract: The Sixteenth International Conference on Knowledge Organization was to have been held in Aalborg, Denmark in July 2020. Cancelled due to COVID-19, the proceedings were published online on 4 December 2020 containing 48 full papers, 17 short papers and 14 posters. Informetric analysis of the proceedings reveals the shifting intension and extension of the knowledge organization domain. International participation was extensive as usual. There is a much larger share of empirical and applied technical research, and therefore much less historical or analytical work than before. The shape of the research front continues to revolve around concept theory and domain analysis, but cultural and ethical issues are more prevalent than before, having attained nearly core status. There is new emphasis on concepts around interdisciplinarity and phenomenon-based knowledge organization systems, and facet analytical theory has been extended into new approaches in the linked data environment. There are more journal articles than before. No monographs are highly cited but interdisciplinarity, Wittgenstein, domain analysis and music classification are prominent alongside archival science and ontology construction on the bookshelf for this conference. Among conferences ISKO and its chapters predominate. The discourse represented by the works on ISKO 16’s bookshelf seems to be rooted in classical concept theory. There is some concretization of discourse concerning interoperability and the continued distancing from ideas of single or “universal” knowledge organization systems. There is a recurrence of concepts of warrant buttressing the extension of research into issues of culture and identity.

Keywords: ISKO international conference, knowledge organization, citations, references

1.0 The Year 2020

It is safe to say none of us knew. None of us knew when the new year dawned on 1 January 2020 what sort of year we would be living through. None of us who attended ISKO 15 in Porto knew that the 2020 conference planned for Aalborg would not take place. When the first news of COVID-19 came, none of us knew it would be a life-changing entity. And none of us knew how a virus could change the course of our science. Certainly, none of us knew that not only would we not gather in 2020 but that the world would have to wait for a vaccine to find some semblance of order on the lines of what we had known before. But, as we shall see, the science of knowledge organization is robust and evolving. The challenge for the science of knowledge organization is to concretize now, to pull up the drawbridges of distraction, to enhance the core membership, the strengthen the theoretical tenets of the domain, and to do all of this during a pandemic.

The Sixteenth International Conference on Knowledge Organization was to have been held in Aalborg, Denmark in the northern tip of Europe. A call for papers was issued in mid-2019. A local committee insisted that there must be a two-tier peer-review system of abstracts first and papers
later. The norm for ISKO international conferences for decades has been to elicit new ideas by asking for abstracts, and then having accepted abstracts, to let the authors have free reign to produce papers of value. The method employed on this occasion meant that many authors would refuse to participate, and that many others would drop out as the parameters of inconsistent peer review became apparent.

Following the cancellation of the conference due to COVID-19, ISKO agreed to pay Ergon Verlag to publish the proceedings for online distribution to the membership. The “volume” was published on 4 December 2020 via the Nomos eLibrary (which separates the volume from the collected proceedings in the series Advances in Knowledge Organization that are available to ISKO members on the Ergon Verlag microsite (https://www.ergon-verlag.de/isko_ko/). The volume is titled Knowledge Organization at the Interface: Proceedings of the Sixteenth International ISKO Conference, 2020 Aalborg, Denmark, ed. Marianne Lykke, Tanja Svarre, Mette Skov and Daniel-Martinez-Avila. The volume is no. 17 in the series Advances in Knowledge Organization. According to the “Introduction” (11) conference topics were:

Knowledge organization across domains, media and technologies
Knowledge organization as understanding and communication
Knowledge organization as driver for development and change.

The volume contains “48 full papers, 17 short papers and 14 posters.” Contents are described as (11-12):

The papers covered a wide range of topics within the conference themes, i.e., knowledge transfer, concepts and conceptualization, fiction genres, ethical aspects, classificatory structures, representation, probabilistic models, social tagging, domain analysis, music classification, legacy data, document types, semantic networks, bibliographic relationships, faceted classification, KOS mapping, warrants, KO education, museum
organization, and archival organization. The papers discussed theoretical issues related to knowledge organization and the design, development and implementation of knowledge organizing systems as well as practical considerations and solutions in the application of knowledge organization theory. The papers covered knowledge organization systems ranging from classification systems, thesauri, metadata schemas through to ontologies and taxonomies.

Contents are “organized in full papers, short papers and posters ... alphabetically by first author.” Since 2008 I have analyzed the contents of international ISKO conferences in a series of editorials titled “ISKO’s Bookshelf” (Smiraglia 2008, 2011, 2013, 2014, 2016, 2018). Following on the concept of authors as citers over time (White 2001) and a 2015 refinement of the notion of a single scholar’s domain (Smiraglia and Beak 2015) my approach has been to discover through informetric analysis of the works cited by the community the following:

the shifting intension and extension of the domain of knowledge organization (KO),
the international growth of the domain
the influential discourses that were motivating ISKO scholars
signs of concretization of the science of knowledge organization.

Research questions across the series have remained:

– How international is participation in ISKO’s international conferences and where are global centers of innovation located?;
– What epistemic stances of ISKO scholars are visible in the citation indicators at each international conference?;
– What is the shape of the research front as represented by each international conference?; and,
– How are the shifting extension and intension of the KO domain reflected in each international conference?

The proceedings volume contain 79 entries, as described above: 48 full papers, 17 short papers and 14 poster. (The volume offered to members by Nomos/Ergon was not actually available on 4 December 2020—logins provided by the publisher to conference contributors were not functional. This analysis was conducted from a draft provided by the series editor.) The number of collaborative papers has increased even since the remarkable 68.6% reported in 2018—fully 75.9% of contributions in 2020 came from collaborative teams ranging in size from two to eight. Half were by two authors, 27% were by three authors, and the remaining 20% were from teams of four to eight authors. This result shows a subtle shift in the domain from its more or less equal division among empirical and humanistic epistemologies toward the characteristics of a more empirical science.

2.0 Pandemic geopolitics

ISKO biennial conferences have always been international. Over time it has been interesting to monitor the growth of the domain in different geopolitical environments. By analyzing the reported institutional affiliations of authors we
can create a visualization of the geopolitical origins of the contents of the proceedings. This year’s visualization appears in Figure 1.

Although the proportions are different from in 2018 (Smiraglia 2018 345) Brazil, the United States and Canada are still the largest contributors; Brazil’s contribution is smaller and the US and Canadian contributions are larger than 2018. There are fewer countries of affiliation this time, and there seem to be fewer contributions from Asia. According to the unpublished report to the 2020 ISKO General Assembly (Smiraglia 2020), the majority of manuscript submissions to the journal come from China (18%), Brazil (12%), the United States (10%), Canada (7%) and the United Kingdom (6%). Interestingly, half of the poster submissions (originated in institutions with Brazilian affiliations. An unanswerable question is to what extent the encroaching pandemic affected submission of abstracts as the pandemic took hold first in China in late 2019 and then spread to Italy and then to North American in early 2020.

Collaborative teams were analyzed by country of affiliation; the results appear in Table 1.

The largest collaborative teams came from South Korea, Austria, Mexico and Canada. Smaller collaborative teams were associated with institutions in Brazil, the USA and Canada—the countries of affiliation contributing the largest proportion of content.

<table>
<thead>
<tr>
<th>affiliation</th>
<th>collaboration</th>
</tr>
</thead>
<tbody>
<tr>
<td>South Korea</td>
<td>6</td>
</tr>
<tr>
<td>Austria</td>
<td>5</td>
</tr>
<tr>
<td>Mexico</td>
<td>4</td>
</tr>
<tr>
<td>Canada</td>
<td>3.5</td>
</tr>
<tr>
<td>China</td>
<td>3</td>
</tr>
<tr>
<td>Nigeria</td>
<td>3</td>
</tr>
<tr>
<td>Portugal</td>
<td>2.7</td>
</tr>
<tr>
<td>UK</td>
<td>2.7</td>
</tr>
<tr>
<td>Spain</td>
<td>2.6</td>
</tr>
<tr>
<td>Brazil</td>
<td>2.4</td>
</tr>
<tr>
<td>Algeria</td>
<td>2</td>
</tr>
<tr>
<td>Denmark</td>
<td>2</td>
</tr>
<tr>
<td>France</td>
<td>2</td>
</tr>
<tr>
<td>Germany</td>
<td>2</td>
</tr>
<tr>
<td>Iran</td>
<td>2</td>
</tr>
<tr>
<td>Singapore</td>
<td>2</td>
</tr>
<tr>
<td>Switzerland</td>
<td>2</td>
</tr>
<tr>
<td>Uruguay</td>
<td>2</td>
</tr>
<tr>
<td>Italy</td>
<td>1.7</td>
</tr>
<tr>
<td>USA</td>
<td>1.5</td>
</tr>
<tr>
<td>Australia</td>
<td>1</td>
</tr>
<tr>
<td>Cameroon</td>
<td>1</td>
</tr>
<tr>
<td>Norway</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 1. Collaborations by Country
3.0 Citing behavior of ISKO 16 authors

There were 1241 citations to source works cited by the authors of the 79 contributions. The mean number of references per contribution was 15.7. The mean number of references per paper was 17.7 (range 2 to 47), per poster 6.1 (range 1 to 12). The mean age of work cited was 14.8 years (14.7 for papers and 15.1 for posters). These figures are consistent with prior conferences, suggesting the continued social scientific orientation. Age of work cited ranged from 2 to 107. Plotting mean age of work cited by authors from different regions yielded the distribution in Table 2.

Brazil, the USA and Canada again fall into the middle of the distribution suggesting the majority of the papers and posters fall along the mean in terms of age of cited work.

Typically in KO the most cited sources are articles from refereed journals. Conference papers are also an important source of recent science, but monographs, technical reports and chapters from anthologies also constitute a large share of source material. Table 3 shows the distribution of publication venues; the category “monographs, etc.” includes technical reports and chapters from anthologies.

This particular conference seems to have a slightly more scientific bent than usual, with a much larger proportion (approximately twice as much) of journal articles and conference papers than other sources.

Fifty-five journals were cited two or more times. A very eclectic list of 200 journals were cited once; these range from the *Wiener Zeitschrift für Die Kunde Südasiens* to *Journal of the History of Sexuality* to *Journal of Information Warfare* to *Chicago Review, Art & The Public Sphere* and *Tourism Geographies*. The mean age of work cited in journal articles is 11.5 years. Table 4 is a list of the 21 journals cited four times or more.

This is a shorter list of journals in the top tier than we have seen in earlier conferences. *Knowledge Organization* is always the journal most cited by ISKO conference authors but the proportionate share of citations to *KO* is much larger than has been seen before.

Specific articles that receive recurring citation point to the discourse informing the conference domain. The list of journal articles cited two or more times appears in Table 5.

<table>
<thead>
<tr>
<th>affiliation</th>
<th>mean age of work cited</th>
</tr>
</thead>
<tbody>
<tr>
<td>Norway</td>
<td>28.7</td>
</tr>
<tr>
<td>Italy</td>
<td>23.7</td>
</tr>
<tr>
<td>Austria</td>
<td>22</td>
</tr>
<tr>
<td>Algeria, USA</td>
<td>20.7</td>
</tr>
<tr>
<td>Uruguay</td>
<td>17.9</td>
</tr>
<tr>
<td>Brazil</td>
<td>17</td>
</tr>
<tr>
<td>Spain</td>
<td>16.9</td>
</tr>
<tr>
<td>Mexico</td>
<td>15.7</td>
</tr>
<tr>
<td>Cameroon</td>
<td>14</td>
</tr>
<tr>
<td>USA</td>
<td>13.3</td>
</tr>
<tr>
<td>UK</td>
<td>13</td>
</tr>
<tr>
<td>Canada</td>
<td>12.1</td>
</tr>
<tr>
<td>Singapore</td>
<td>11.8</td>
</tr>
<tr>
<td>France</td>
<td>10.7</td>
</tr>
<tr>
<td>Germany</td>
<td>10.2</td>
</tr>
<tr>
<td>Portugal</td>
<td>9.7</td>
</tr>
<tr>
<td>Australia</td>
<td>7.7</td>
</tr>
<tr>
<td>Denmark</td>
<td>7</td>
</tr>
<tr>
<td>China</td>
<td>6.5</td>
</tr>
<tr>
<td>Iran</td>
<td>6.2</td>
</tr>
<tr>
<td>Nigeria</td>
<td>5.8</td>
</tr>
<tr>
<td>South Korea</td>
<td>5.5</td>
</tr>
<tr>
<td>Austria</td>
<td>5.2</td>
</tr>
<tr>
<td>Switzerland</td>
<td>2.5</td>
</tr>
</tbody>
</table>

*Table 2. Mean age of work cited by affiliation*

<table>
<thead>
<tr>
<th>Venue</th>
<th>Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Journal articles</td>
<td>602</td>
</tr>
<tr>
<td>Conference papers</td>
<td>198</td>
</tr>
<tr>
<td>Monographs etc.</td>
<td>403</td>
</tr>
<tr>
<td>Dissertations and theses</td>
<td>17</td>
</tr>
<tr>
<td>Web resources</td>
<td>68</td>
</tr>
</tbody>
</table>

*Table 3. Publication venues.*
Table 4. Most cited journals.

<table>
<thead>
<tr>
<th>Title</th>
<th>Year</th>
<th>Citation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Knowledge Organization</td>
<td>137</td>
<td></td>
</tr>
<tr>
<td>Cataloging &amp; Classification Quarterly</td>
<td>33</td>
<td></td>
</tr>
<tr>
<td>Data Science Journal</td>
<td>32</td>
<td></td>
</tr>
<tr>
<td>Journal of Statistical Software</td>
<td>32</td>
<td></td>
</tr>
<tr>
<td>Journal of the Association for Information Science and Technology</td>
<td>32</td>
<td></td>
</tr>
<tr>
<td>Journal of the Korean Society for Information Management</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>Library Trends</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Novos Estudos</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Journal of the American Society for Information Science &amp; Technology</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>Notes</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>International Journal on Digital Libraries</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>Bulletin of the American Society for Information Science and Technology</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Journal of Risk Research</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Library Resources &amp; Technical Services</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Archives and Museum Informatics</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Axiomathesis</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Journal of Travel Research</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>American Archivist</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Information Research</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>International cataloguing and bibliographic control</td>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>

[Table 4. Most cited journals.]

<table>
<thead>
<tr>
<th>Author(s)</th>
<th>Year</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beghtol, Clare</td>
<td>2002</td>
<td>&quot;A Proposed Ethical Warrant for Global Knowledge Representation and Organization Systems&quot;</td>
</tr>
<tr>
<td>Giunchiglia, Fausto, Biswanath Dutta, and Vincenzo Maltese</td>
<td>2014</td>
<td>&quot;From Knowledge Organization to Knowledge Representation&quot;</td>
</tr>
<tr>
<td>Gnoli, Claudio</td>
<td>2016</td>
<td>&quot;Classifying Phenomena Part 1: Dimensions&quot;</td>
</tr>
<tr>
<td>Hudon, Michele</td>
<td>1997</td>
<td>&quot;Multilingual Thesaurus Construction-Integrating the Views of Different Cultures in One Gateway to Knowledge and Concepts&quot;</td>
</tr>
<tr>
<td>Zeng, Lei</td>
<td>2019</td>
<td>&quot;Interoperability&quot;</td>
</tr>
<tr>
<td>Abgaz, Yalemisew, Amelie Dorn, Barbara Piringer, Eveline Wandl-Vogt, and Andy Way</td>
<td>2018</td>
<td>&quot;Semantic Modelling and Publishing of Traditional Data Collection Questionnaires and Answers&quot;</td>
</tr>
<tr>
<td>Alajmi, Bibi and Sajjad Rehman</td>
<td>2016</td>
<td>&quot;Knowledge Organization Trends In Library And Information Education: Assessment And Analysis&quot;</td>
</tr>
<tr>
<td>Barić, Mario</td>
<td>2019</td>
<td>&quot;Towards a General Concept of Warrants: First Notes&quot;</td>
</tr>
<tr>
<td>Bates, Jo and Jennifer Rowley</td>
<td>2011</td>
<td>&quot;Social Reproduction and Exclusion in Subject Indexing: A Comparison of Public Library OPACs and LibraryThing Folksonomy&quot;</td>
</tr>
<tr>
<td>Adler, Melissa</td>
<td>2009</td>
<td>&quot;Transcending Library Catalogs: A Comparative Study of Controlled Terms in LCSH and User-generated Tags in LibraryThing for Transgender Books&quot;</td>
</tr>
<tr>
<td>Barić, Mario</td>
<td>2018</td>
<td>&quot;Literary Warrant&quot;</td>
</tr>
<tr>
<td>Beghtol, Clare</td>
<td>1986</td>
<td>&quot;Semantic Validity: Concepts of Warrant in Bibliographic Classification Systems&quot;</td>
</tr>
</tbody>
</table>
Table 5. Most cited journal articles.

<table>
<thead>
<tr>
<th>Author(s)</th>
<th>Year</th>
<th>Title</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Broughton, Vanda</td>
<td>2006</td>
<td>“The Need for a Faceted Classification as the Basis of All Methods of Information Retrieval”</td>
<td>2</td>
</tr>
<tr>
<td>Buckland, Michael K</td>
<td>1997</td>
<td>“What is a ‘Document’?”</td>
<td>2</td>
</tr>
<tr>
<td>Bullard, Julia</td>
<td>2016</td>
<td>“Warrant as a Means to Study Classification System Design”</td>
<td>2</td>
</tr>
<tr>
<td>Chen, Peter Pin-Shan</td>
<td>1976</td>
<td>“The Entity-Relationship Model—Toward a Unified View of Data”</td>
<td>2</td>
</tr>
<tr>
<td>Colombo, Stephanie and Mario Barité</td>
<td>2015</td>
<td>“Tres Enfoques de Bias en Organización del Conoci-miento: Bias Neutro, Bias Negativo y Bias Positivo”</td>
<td>2</td>
</tr>
<tr>
<td>Duranti, Luciana</td>
<td>1997</td>
<td>“The Archival Bond”</td>
<td>2</td>
</tr>
<tr>
<td>Gnoli, Claudio</td>
<td>2018</td>
<td>“Classifying Phenomena Part 4: Themes and Rheemes”</td>
<td>2</td>
</tr>
<tr>
<td>Golder, Scott A and Bernardo A Huberman</td>
<td>2006</td>
<td>“Usage Patterns of Collaborative Tagging Systems”</td>
<td>2</td>
</tr>
<tr>
<td>Hider, Philip</td>
<td>2015</td>
<td>“A Survey of the Coverage and Methodologies of Schemas and Vocabularies Used to Describe Information Resources”</td>
<td>2</td>
</tr>
<tr>
<td>Hjørland, Birger and Hanne Albrechtsen</td>
<td>1995</td>
<td>“Toward a New Horizon in Information Science: Domain-Analysis”</td>
<td>2</td>
</tr>
<tr>
<td>Hjørland, Birger</td>
<td>2017</td>
<td>“Domain Analysis”</td>
<td>2</td>
</tr>
<tr>
<td>Lee, Deborah and Lyn Robinson</td>
<td>2017</td>
<td>“The Heart of Music Classification: Toward a Model of Classifying Musical Medium”</td>
<td>2</td>
</tr>
<tr>
<td>Lee, Deborah</td>
<td>2017</td>
<td>“Numbers, Instruments and Hands: The Impact of Faceted Analytical Theory on Classifying Music Ensembles”</td>
<td>2</td>
</tr>
<tr>
<td>Lykke Nielsen, Marianne</td>
<td>2001</td>
<td>“A Framework for Work Task Based Thesaurus Design”</td>
<td>2</td>
</tr>
<tr>
<td>Mazzocchi, Fulvio</td>
<td>2018</td>
<td>“Knowledge organization system (KOS)”</td>
<td>2</td>
</tr>
<tr>
<td>Noruzi,</td>
<td>2012</td>
<td>“FRBR and Tillett’s”</td>
<td>2</td>
</tr>
<tr>
<td>Alireza</td>
<td>2001</td>
<td>“Constructing a Function-Based Records Classification System: Business Activity Structure Classification System”</td>
<td>Archivaria 51: 137-154</td>
</tr>
<tr>
<td>Sabourin, Paul</td>
<td>2013</td>
<td>“Metaphors of Time and Installed Knowledge Organization Systems: Ouroboros, Architectonics, or Lachesis?”</td>
<td>2</td>
</tr>
</tbody>
</table>

Of course, only partial citations are given here; full citations appear in the proceedings. The table is arranged in order of descending frequency of citation. The comparable list from ISKO 15 had 27 authors; there are 40 on this list, suggesting a broader source of discourse. On the other hand, the articles most frequently cited are cited more often than before, suggesting a more concrete reliance on those articles for discourse. As usual, Hjørland and Dahlberg are the most heavily cited, followed by Tennis, Beghtol and Gnoli, but this time newcomers Adler, Barité and Guinchiglia Dutta and Maltese join this center tier together with KO regulars Hjørland and Zeng. We also can note that there is some ritual citation here of Hjørland’s outdated 2002 article; the cumulative nature of science would suggest there should be higher citation of his 2017 update. In addition to signaling ritual citation (see also the citation to Hjørland and Albrechtsen 1995) this also is a sign of loose refereeing. We see here the core of knowledge organization (concept theory and classification) alongside warrant (or meaning) and unsettled factions of applicability ranging from interoperability to conceptual models. Probably the clearest discourse element in this table is the search for a unifying conceptual model. The category “Monographs, etc.” in Table 3 includes monographs, chapters in anthologies and technical reports or government documents. There were 219 citations to
monographs, 119 to chapters in anthologies, and 68 to technical reports or government documents. Eight technical reports or government documents were cited more than once; these appear in Table 6.

Thirteen monographs were cited more than once. These appear in Table 7.

Four chapters (Green, Khoo, Gnoli and Szostak) were cited more than once; five anthologies were cited more than once. The emergence of the online ISKO Encyclopedia of Knowledge Organization is notable, in particular if we consider its relationship in KO to the role of the former Annual Review (ARIST) in information. Both chapters and anthologies are included in Table 8.

Conference proceedings are almost as important in KO as journal articles. This is a sign of the activity level of a relatively new science, where there is much new work underway based on a slender theoretical core. Conference proceedings most cited appear in Table 9. ISKO 11 (Rome 2010) was cited 6 times, ISKO Brazil 2019 and UDCC London 2017 each 5 times, ASIST 53 2016 4 times.

3.0 The research front

A usual first step in identifying a domain’s research front is identifying the authors whose work is most cited. It is common to see a mixture of ritual citation (e.g., Ranganathan, cited in nearly every paper that concerns “facets”) and citation to those whose work has most influenced the authors contributing current research to the conference. 153 authors were cited twice or more; 28 were cited 5 times or more—these are shown in Table 10.


Table 6. Technical or government reports.

| Szostak, Rick, Claudio Gnoli, and María José López-Huertas. 2016. Interdisciplinary Knowledge Organization. Switzerland: Springer. | 3 |
| Cabré, María Teresa. 1999. La Terminología: Representación y Comunicación: Elementos Para una Teoría de Base Comunicativa y Otros Artículos. Barcelona: Institut Universitari de Lingüística Aplicada. | 2 |
| Olson, Hope A. 2002. The Power to Name: Locating the Limits of Subject Representation in Libraries. Canada: Springer. | 3 |

Table 7. Most cited monographs.


Table 8. Anthologies and chapters with multiple citations.

<table>
<thead>
<tr>
<th>Conference</th>
<th>Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ISKO International</td>
<td>43</td>
</tr>
<tr>
<td>ISKO Chapters</td>
<td>21</td>
</tr>
<tr>
<td>ASIST</td>
<td>9</td>
</tr>
<tr>
<td>UDCC</td>
<td>5</td>
</tr>
<tr>
<td>Other</td>
<td>27</td>
</tr>
</tbody>
</table>

Table 9. Most cited conferences.

<table>
<thead>
<tr>
<th>Author</th>
<th>Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hjørland</td>
<td>41</td>
</tr>
<tr>
<td>Smiraglia</td>
<td>19</td>
</tr>
<tr>
<td>Gnoli</td>
<td>17</td>
</tr>
<tr>
<td>Szostak</td>
<td>17</td>
</tr>
<tr>
<td>Dahlberg</td>
<td>14</td>
</tr>
<tr>
<td>Beghtol</td>
<td>13</td>
</tr>
<tr>
<td>Guimaraes</td>
<td>10</td>
</tr>
<tr>
<td>Lee, Deborah</td>
<td>10</td>
</tr>
<tr>
<td>Olson</td>
<td>10</td>
</tr>
<tr>
<td>Ranganathan</td>
<td>9</td>
</tr>
<tr>
<td>Tennis</td>
<td>9</td>
</tr>
<tr>
<td>Adler</td>
<td>8</td>
</tr>
</tbody>
</table>

Table 10. Most cited authors.
Author co-citation analysis is a technique that helps uncover theoretical poles in a domain’s discourse by correlating perceptions among citing authors of relationships among research from key members of the research front. It is important to remember that the visualization represents the perceptions of citing authors, in this case it represents the perceptions of authors who contributed to ISKO 16. Thus the perceptions represented in the visualization reveal epistemological rather than ontological connections—a bit of insight therefore into the discourse governing the generation of the research reported by conference submissions. As before (see for example Smiraglia 2018) there was very little co-citation, which is a sign of little theoretical coherence across the conference. Authors from Table 10 were entered into a matrix and co-citations to their work were recorded. Authors with little or no co-citation were then removed, producing a final matrix of 16 authors. A visualization produced using Gephi 0.9.2 (https://gephi.org/) appears in Figure 1.

Again, as in prior conferences, we see the core importance of Dahlberg and Hjørland, and following the thickness of the line we see that they are closely related to Green and Smiraglia. Emanating from Dahlberg is a theoretical region representing aspects of concept theory. Emanating from Hjørland is a theoretical region, including domain analysis. Nodes in that region are represented by Smiraglia and Guimarães. A strong association then emerges between this core and the leading edge authors Szostak, Tennis and Zeng. Szostak (through connections with Kleineberg and Gnoli) represents new approaches to phenomenon-based classification. Tennis Adler and Olson anchor a large region representing ethical and cultural considerations. New names emerging from this visualization are Barité, Adler, García Gutiérrez and Kleineberg. Also interesting is the reemergence of Beghtol (cultural warrant, ethics) and Hudon (multilingual thesauri), whose earlier work is being revisited.

Figure 1. Author co-citation visualization.
4.0 Explicit themes of ISKO 16

Because the conference was cancelled we have no program to work from—typically these include panel titles that provide broad themes. For the present analysis we turned to the titles and abstracts of conference submissions. These were analyzed using Provalis ProSuite’s QDA Miner and WordStat modules (https://provalisresearch.com/products/prosuite-text-analytics-tools/). Among the titles there were 945 words of which 389 were unique; of the 556 recurring terms 19 occurred 3 or more times; these are plotted using WordStat’s MDS visualization tool in Figure 2.

We see here pointers to social and cultural issues, to archival and library systems, to core ideas of domain analysis, ontology, knowledge organization systems, and concepts. To better visualize thematic concepts we analyzed abstracts as well. The abstracts used 12226 words of which 2434 were unique and thus 9792 words recur. To visualize thematic content we used WordStat’s phrase-finder module. 28 phrases of 2 to 5 words occur four times or more; an MDS plot appears in Figure 3.

We see that the recent uptick in emphasis on archival science and knowledge organization constitutes a segment, as does the relationship between faceted classification and information retrieval. Linked data and mapping issues occupy a region that also includes performing arts (not visible is “semantic web”). Finally is the region at the left containing knowledge representation, and semantic relations (behind “extraction of semantic relations” is “semantic relations” alone). We can see from these two visualizations how the core of the domain of knowledge organization is being extended to embrace new linked data and semantic relations technologies.

5.0 ISKO 16’s Bookshelf

Before we ask what is on ISKO’s bookshelf, we might return to the original research questions. Concerning international participation we see this extended as usual. Submissions from Brazil and the United States continue to dominate but many more countries are represented as well. A continuing anomaly is the low participation of Chinese contributors given the large number of Chinese submissions to the society’s Knowledge Organization journal. Concerning epistemic stances, we see in this particular conference a much larger share of empirical and applied technical research, and therefore much less historical or analytical work than before.
The shape of the research front continues to revolve around concept theory and domain analysis, but we see that cultural and ethical issues are more prevalent than before, having attained nearly core status. We also see new emphasis on concepts around interdisciplinarity and phenomenon-based knowledge organization systems. The utility of facet analytical theory has been extended from its original use in classification into new approaches to knowledge representation in the linked data environment of the semantic web. Thus the extension continues to shift related to the influence of information technology as the intension solidifies around core concepts, which now include cultural and ethical concerns.

What is on ISKO 16’s bookshelf? There are more journal articles than before, and a greater proportion of those are from the society’s Knowledge Organization. But there is an expanding list of contributing journals as well. Thesauri and the new Library Reference Model are prominent. No monographs are highly cited but interdisciplinarity, Wittgenstein, domain analysis and music classification are prominent alongside archival science and ontology construction on the bookshelf for this conference. Among conferences ISKO and its chapters predominate.

The discourse represented by the works on ISKO 16’s bookshelf seems to be rooted in classical concept theory, teased into new territory by the notions of interdisciplinarity and freedom from hierarchical constructs represented by phenomenon-based approaches and the extension of facet analytical theory. Domain analysis shows the concretization of discourse concerning interoperability and the continued distancing from ideas of single or “universal” knowledge organization systems. Perhaps the most obvious element of discourse in ISKO 16 is the recurrence of warrant buttressing the extension of research into issues of culture and identity. Thus the discourse remains robust. ISKO is well placed to continue to make a difference in a post-pandemic world.

References


Reality, Unreality, and Artistic Deception: The Ethical Dimensions of the Knowledge Organization of Art Documentation†

Deborah Lee

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

Abstract: The ethics of describing and indexing works which have an element of deception is an important topic within the arena of knowledge organization (KO). However, what happens if the unreal element is for artistic purposes and is part of the experience of the document? The focus of this article is on the KO of art documentation. It considers art documents where unreality and mistruth are part of a document’s creative and artistic purpose. Three examples of exhibition documentation for contemporary art are explored. These illuminate the KO implications of unreal elements in artworks, exhibitions and documents, and the interplay between them. Three models are then presented relating to the KO of unreality. Model one shows how the wishes of the creator are a significant part of ethical KO for art and the ethical KO decisions which result from the performative and experiential aspects of art documentation. Model two illustrates how the placement of the unreality in relation to metadata creation can have a critical impact on ethical decisions in KO. Model three posits unreal-ness as a novel type of information about metadata creation, introducing categories such as motive for the deception. It shows how this original way of contemplating applied ethics in KO is vital for all ethically challenging works. Ultimately, considering the KO ethics of art documentation extends our thinking about how to deal with deception and unreality and adds an important aesthetic-ethical dimension to the corpus of work on KO ethics.

Keywords: unreal, works, documentation, art, exhibition, ethical

† The author would like to thank the following: 1) Prof. David Bawden for his advice about this article; 2) the Book Library and library colleagues at the Courtauld Institute of Art for providing inspiration for the examples analyzed in this article; and, 3) the anonymous reviewers for Knowledge Organization for their insightful feedback.

1.0 Introduction

Knowledge organization (KO) discourse about ethical responsibility shows how the act of classifying and cataloging a document has an ethical dimension (for example, Bair 2005; Shoemaker 2015; Snow 2015). An important aspect of this research has been to consider the ethics of dealing with works that are in some way deceptive. This is particularly pertinent considering the current era of misinformation, alternative facts and untruths. However, creative works frequently play with the idea of truth and reality, and these artistic untruths and unreali
siders the ethical considerations for art documentation with elements of the unreal and asks how such works can expand and challenge existing thinking about ethical KO.

Exhibition catalogs are, at their barest, works that document and record an exhibition including its artworks and artists (ARLIS/NA 2010; Grandal Montero 2017); however, the late twentieth-century and twenty-first-century exhibition catalog will frequently include much more, such as critical essays, biographical information, general surveys of the artists’ works and analyses of the exhibited artworks (ARLIS/NA 2010; Grandal Montero 2017). Typically, art exhibition catalogs are created by curators and others working on behalf of the museums or galleries responsible for the exhibition. Textual content for the catalogs is produced by the curators themselves and writers commissioned by those curators or catalog editors, which for exhibitions relating to contemporary works is also likely to include the artists. While an exhibition catalog’s purpose is to document the exhibition and artworks, experimental formats of exhibition catalog will frequently include much more, such as critical essays, biographical information, general surveys of the exhibited artworks (ARLIS/NA 2010; Grandal Montero 2017). The resulting blurred lines between artwork and documentation have ramifications for cataloging and ethics and are a central tenet of this article. It should be noted that not all works published in conjunction with an exhibition explicitly describe themselves as catalogs or even include a list of artworks. Therefore, the broader term “exhibition documentation” is utilized in this article, following the example set by ARLIS/UK & Ireland’s (2000) important cataloging guidelines; so, “exhibition documentation” is used as a broad category that refers to works that in some way document exhibitions or other limited-time display of artworks, including those that are unquestionably exhibition catalogs.

The methodology employed in this article is three-fold. Literature is analyzed in order to utilize existing findings born out of related scenarios, such as false memoirs (Snow 2015). Three examples of exhibition documentation are analyzed, using a framework built from ideas in Snow (2015). The three examples were selected as they are known to the author for containing different examples of creative unreality. While the examples illustrate a variety of scenarios and, taken together, help to illuminate the parameters of this intriguing type of work, the three examples are not intended to be representative of artistic unreal works. Models are created from the literature analysis and from ideas elicited from the three examples of exhibition documentation. The article has a number of purposes and limitations. Although KO is a broad set of practices, this article is chiefly concerned with bibliographic description and indexing (The terms resource description, metadata creation and cataloging are used interchangeably throughout the article, and indexing refers to the assignment of index terms from subject indexes and thesauri). Significantly, this article does not provide detailed catalog records of the three example artworks, nor state the preferred way to approach these items. Instead, it is chiefly concerned with the broad ethical outlook that the metadata creator is faced with before commencing description and indexing and the consequences of decisions surrounding the ethics/aesthetics boundary. So, arguably, this article is about meta-metadata. It is also acknowledged that the scope of artistic creativity contemplated in this article is narrow in a number of ways: 1) it is based on art rather than the arts more generally; 2) it relates to one particular type of art documentation that is termed exhibition documentation, and the examples stem from exhibitions and installations; 3) it focuses on three relatively contemporary examples of which two relate to North Korea.

This article starts by analyzing the KO ethics literature that pertains specifically to applying knowledge organization systems and creating descriptions. It identifies two key works in this area by Brubaker (2018) and Snow (2015), and the latter is used as an analysis tool for artistic works. Then, three examples of documents relating to art exhibitions are dissected, each one highlighting how creative works resonate with, and challenge, discussions about metadata for other types of deceptive works. The article culminates in three models that consider creativity and deepen our understanding about unreality in applied KO. Ultimately, some interesting questions arise when art meets ethics in KO, and this enriches our general understanding of how we should deal with documents of deception, mistruth and unreality.

2.0 Literature analysis of applied ethics and deceptive works

2.1 Introducing applied ethics literature in knowledge organization

Ethics forms an important part of research within KO. However, despite this, Fox and Reece (2012, 378) suggest that the precise nature of what is meant by ethics is not frequently discussed, and more often relies on a “sense of right and wrong not rigorously defined.” As the research in this article is reliant on the existing corpus of KO ethics literature, Fox and Reece’s (2012) depiction of a loose and flexible meaning of ethics will be adopted; so, in this article, the term ethics will be used very generally to mean discussions that evoke morality in its broadest sense, without discussions about how this fits into the branch of philosophy or its associated principles.

A range of topics and perspectives make up ethics KO discourse. Some of the discourse could be categorized as being focused on particular groups of people, such as gender (for example, Olson 2007; Fox 2014; Adler 2016). However, there is also often a separation in the literature between the
following: 1) discussions about ethical issues in knowledge organization systems (KOSs); and, 2) works that consider the processes of constructing metadata for documents, including the application of those KOSs to individual documents. While both concern applied ethics, the second category is of most relevance to this article. The applied ethics of constructing metadata includes a variety of different approaches. For example, some works focus on codes and guidelines for doing ethical KO in practice (for instance, Shoemaker 2015; Bair 2005); others focus on specific scenarios or types of material, such as ethical issues in authority control for zine authors (Fox and Swickard 2019) or dealing with so-called “bad books” (Homan 2012). Therefore, this article utilizes the KO applied ethics literature, which discusses ethical codes and guidelines, yet uses it to discuss a very specific scenario involving art documentation. Literature that discusses so-called deceptive works, provides the foundations for analyzing the novel addition of aesthetics to the ethical conundrums. Therefore, the KO literature analysis focuses on works within two areas: the development of ethical codes of practice in the Anglo-American cataloging communities and KO for deceptive works.

2.2 Ethical codes of practice and ethical guidelines

Bair’s (2005) seminal article discusses the principles pertaining to ethical responsibilities for catalogers and she introduces a code of ethics specifically for catalogers. However, this code was not adopted by the community; Shoemaker (2015) suggests that one reason for this is that it is the creation of an individual rather than a collection of librarians. Fox and Reece (2012) also present a set of ethical guidelines, which examines what is meant by morality in KO. Their guidelines include the duty to care about what happens, the idea of being rigid but with some mitigation and the importance of considering the consequences of KO decisions (Fox and Reece 2012). The final point here may be especially pertinent to cataloging works about art, where the documents accompanying artworks can be part of the aesthetic experience of the artwork and there may be corresponding consequences of including or not including information in the metadata. Ultimately, Fox and Reece’s (2012) guidelines relate more to classification than description, so will not be used further in this article.

Snow (2015) offers an alternative to an ethical code of practice. In Snow’s (2015) analysis and discussion about the ethical dilemmas presented by false memoirs, which is discussed in detail below, she utilizes a set of ethical principles that are extracted from the IFLA Statement of International Cataloging Principles (ICP). These statements, produced by the cataloging section of IFLA (Galeffi et al. 2017), are originally based on the 1961 Paris principles of cataloging. The ICP was first published in 2009, which is the version used in Snow (2015). The most recent version will be used in this article and dates from 2017 (There are some differences between the 2009 and 2017 versions, in the number of principles and the short descriptions of the principles. However, the differences are not substantive for the topic of discussion in this article). Snow (2015) draws upon three of these principles in her discussion: 2.1 Convenience of User, 2.3 Representation and 2.4 Accuracy. In the ICP (Galeffi et al. 2017), the first principle is given as the most important and if necessary, the one that should be prioritized (Galeffi et al. 2017): “2.1 Convenience of the user. Convenience means that all efforts should be made to keep all data comprehensible and suitable for the users.” This means that from an ethical perspective, the users’ needs should be prioritized when faced with difficult cataloging decisions. The other two principles discussed by Snow (2015) are as follows: “2.3 Representation. A description should represent a resource as it appears” (Galeffi et al. 2017) and “2.4 Accuracy. Bibliographic and authority data should be an accurate portrayal of the entity described.” Note that the full description for ICP 2.3 discusses authority control and controlled names, although it can be appropriated to cover all aspects of cataloging including the descriptive elements as well.

It is important to consider who the ethical obligations of the catalogers are to. Bair (2005) gives a list of six interconnected and overlapping groups: society, institutions that catalogers work at, library users, librarians, the profession and generally to humans. It is interesting to note that the creators of the resources do not feature explicitly in this list, nor in many other discussions about ethical obligations, and this omission will be explored in the discussions about metadata for artistic materials.

Finally, in the last few years, cataloging communities in the U.S., Canada and the U.K. have been collaborating on developing a code of ethics for catalogers, under the guise of the Cataloging Ethics Steering Committee (2020a). In June 2020, the first draft was released, and the second draft is the most current version available at the time of writing (Cataloging Ethics Steering Committee 2020a). As this work is quite broad in nature, there is little that is germane to the detailed ethical decisions about individual metadata elements that form the basis of this article. Two small points are worth mentioning, however. The first of the nine principles says that “we catalogue resources in our collections with the end-user in mind” (Cataloging Ethics Steering Committee 2020b); this could be viewed as correlating with ICP 2.1 (Galeffi et al. 2017), which concerns the convenience of the user, though there is no suggestion in the draft code of ethics, unlike in the ICP, that this principle should trump the other principles. Creators are mentioned in the new code of ethics (Cataloging Ethics Steering Committee 2020b), unlike in the other principles and research discussed so far. However, as this occurs as part of the principle
concerning bias (Cataloging Ethics Steering Committee 2020b), this comment seems to refer to creators as general human beings rather than to their creative vision and is, therefore, not directly relevant to the discussion. So, while there is now a draft of a new code of cataloging ethics, the ICP principles identified and utilized by Snow (2015) are more useful for the analysis in this article. Therefore, ICP 2.1, ICP 2.3 and ICP 2.4 will be used as the framework for discussing art-related ethical questions.

2.3 Deception and deceptive works

A significant and relevant topic in applied ethics in KO concerns the organization and description of works that are “of questionable authorship and authenticity or veracity” (Brubaker 2018, 9). Two sources are particularly useful here: Brubaker’s book of examples of deceptive works (2018) and Snow’s (2015) article analyzing the ethical issues with false memoirs. They are important to the discussion in this article for two main reasons. First, they discuss works that are akin to the art-related examples but, as will be shown, have important differences. Second, the framework developed by Snow (2015) to discuss one specific example of deceptive work can be utilized to interrogate another.

Brubaker’s 2018 book is centered on a set of example problematic works, including memoirs, autobiographies, fictional novels and poetry. Each work is outlined and the author gives her opinion on what solution should (or should not) be adopted. Two aspects of this book are of interest to this article. First, there is discussion about categorizing the deceptions. Brubaker (2018) divides the deceptions into two broad categories: 1) the author misrepresented themselves in some way; and 2) nonfiction works that are either actually fictional or include significant mis-truths. This is useful as it serves as a broad categorization of deceptive works, where the dividing factor is whether the deception centers on the author or the work. It is probably the latter that is most similar to the art documentation considered in this article. Second, Brubaker (2018) very briefly mentions the issue of ethics and creativity, which is the topic of this article. She (Brubaker 2018, 127) describes examples where “deceptions regarding authorship or origin are central to the author’s creative vision, and the reading experience is enhanced by the reader’s growing awareness of the deception,” and suggests there is sometimes a tension between accuracy and art. However, there are a few significant differences between these statements and the focus of this article. The art documentation considered in this article do not generally focus on authorship or origin. Also, in the examples explored in this article, awareness of the deception by the audience is not necessarily a desired outcome.

Snow’s (2015) article focuses on a specific type of deceptive work: the false memoir. This is a work that describes a period of time in an author’s life, but the truthfulness of the resource is questioned in a public way (Snow 2015). Snow (2015) discusses the ethical problems presented by such works, using specific points from the ICP pertaining to ethics (as discussed above). Snow (2015) concludes that for false memoirs, accuracy of what the resource contains (ICP 2.4) might be in opposition to being faithful to how that resource self-represents (ICP 2.3). In other words, accuracy and self-representation cannot always coexist. Snow’s (2015) suggestion is that the answer may lie in representing multiple realities in order to be useful to readers, such as fiction and nonfiction subject headings in the same catalog record (Snow 2015). This insightful discussion can be transformed into a useful framework for the analysis of art documentation: self-representation (ICP 2.3) plus accuracy (ICP 2.4) equals best convenience to the user (ICP 2.1). This is visualized in Figure 1. So, the analysis of the three

![Figure 1](image_url). Using Snow (2015) as a framework to analyze deceptive works.
examples of art documentation will do the following: 1) it will see how accuracy and self-representation interact; 2) it will consider the impact of including multiple realities, including whether this would create convenience to the user; and, 3) perhaps most significantly, the art documentation examples will be mined to see what additional issues and perspectives art documentation adds to metadata creation for deceptive works.

Two other aspects of Snow (2015) are significant for this article. The first concerns notes, which are a typical way to explain controversies on catalog records (Homan 2012; Brubaker 2018). Snow (2015) explains that making these notes requires high skill from catalogers and time, especially in making these notes as non-judgmental as possible. Second, false memoirs open up a question about the temporal dimension: is there an obligation to return to metadata once a deception has been discovered (Snow 2015)? This temporal aspect also needs consideration for art documentation.

The literature analysis has identified the importance of studying applied ethics in KO, in particular when trying to describe and organize works that have an element of deception. However, much of the existing research has focused on examples that are different from artistic scenarios in the reasons for the deceptions. Though deception-for-creativity is identified by Brubaker (2018), there is no detailed discussion of this phenomenon. Therefore, the literature analysis has identified an important gap that this article will start filling: an in-depth exploration and conceptualization of the ethics of KO when the deception is for artistic purposes.

3.0 The ethics of knowledge organization for unreal works about art

3.1 Introducing unreal art documentation

Art documentation contains some intriguing treatments of unreality, and these illuminate issues concerning the ethics of their knowledge organization. On the surface, this art documentation shares similarities with the deceptive works analyzed and categorized by authors such as Snow (2015) and Brubaker (2018). They are works about art, which could in some circumstances be considered artworks in their own right, where what is presented in the published resource is not necessarily a depiction of reality or the truth. Note that unlike genres of works in other artforms such as say novels or plays, art documentation is not traditionally assumed to be unreal in the same way. This is also reflected in their subject cataloging: the Library of Congress Subject Heading form subdivision of “Exhibitions” used for exhibition catalogs, potentially signifies to a library user a different relationship to truth than the form subdivision of “Fiction.” However, although the unreality might be expected, there is a significant difference in intention between art documentation with unreal aspects and, say, false memoirs. In the works under consideration in this article, there is every reason to believe that any intention for misleading the reader is for artistic purposes (The ability to ascertain the intentions of the creators of such works is not straightforward, as will be seen in the examples).

Terminology for these types of works presents a challenge. Brubaker (2018) uses the term “deceptive works” to discuss the broad category of works where there is a mismatch between what is presented and what is real and which covers specific types such as false memoirs. However, the Oxford English Dictionary defines (“Deceive, v., 2a” 1989) the act of deceiving as “To cause to believe what is false; to mislead as to a matter of fact, lead into error, impose upon, delude, ‘take in.’” While some of these explanations could apply to artistic scenarios, other parts of this definition imply that the action is intended for gain or to cause harm. This makes using the term “deceptive works” problematic for these works, and similar terms such as “misleading works” also have a flavor of negativity. Another possibility is to use the term “fiction” or “fictional,” as one definition of fiction in the Oxford English Dictionary is quite appropriate for this scenario (“Fiction, n., 4a” 1989): “The action of ‘feigning’ or inventing imaginary incidents, existences, states of things, etc., whether for the purpose of deception or otherwise.” However, this is problematic due to the term’s additional usage as a specific type of literature (“Fiction, n., 4a” 1989). So, an alternative term, “unreal works,” is used instead. The adjective “unreal” is defined in the Oxford English Dictionary as “Imaginary, illusory, intangible; not real” (“Unreal, adj. (and int.) and n., 1.” 2014) and arguably best encapsulates this aspect of the artistic works while reducing the possibility of misrepresentation. However, it is acknowledged that no term will be a perfect fit and that some of the rejected terms are useful as near-synonyms.

There is a section of discourse in aesthetics and art theory devoted to the idea of “artistic truth” (see for example, Heyl 1950; Price 1949), which asks what it means for an artwork to be true and whether truth is an important quality. This could be seen as a precedent for contemporary discussions about truthiness in art (for example, Armstrong et al. 2012), which consider the blurred edges between reality and unreality and between truth and mistruth. So, while this article focuses on KO issues, it is worth remembering that these discussions about the implications for KO when dealing with the unreal aspects of art documentation, could also be linked to the aesthetics perspective of artistic truth and truthiness.

To examine the ethical considerations presented by cataloging and indexing artistic unreal works, three examples are analyzed. The analysis uses the framework discussed above (see Figure 1), which utilizes Snow’s (2015) findings about
false memoirs. Information about the works and their cataloging is taken from the following sources: the works themselves, secondary information about the works and from the catalog records on the Courtauld Institute of Art Book Library Catalogue (Courtauld Institute of Art 2020). Note that the catalog records were edited or created by the author of this article, so are not a completely independent source.

3.2 Example 1: Die Deutsches Informationsbibliothek Pjöngjang by Sara van der Heide

The first example is exhibition documentation relating to the artist Sara van der Heide and corresponds to her exhibition shown as part of the fifth Guangzhou Triennial, from December 11, 2015 to April 10, 2016 (Heider et al. 2016). The documentation consists of five booklets. Four of these are the exhibition catalog, with each of the four booklets containing the catalog in a different language (English, German, Chinese and Korean). The fifth booklet consists of essays, reflections and images and is entitled “Critical reflections and documentation of The German Library Pjöngyang.”

The documents represent a set of artistic events by Sara van der Heide. The installation is described by the chief curator of the Biennial as a historical reconstruction of the Deutsches Informationsbibliothek Pjöngjang but also as a space for contemplating political and cultural issues (Slager 2016). The real-life but now disbanded Deutsches Informationsbibliothek Pjöngjang was at the Goethe-Institut’s Information Centre Pyongyang and the center was open from June 2004 for five years (Heide et al. 2016). Sara van der Heide describes (Heide et al. 2016, English exhibition catalog, 1) her work as an “intervention ... an imaginary transformation of the current geography of the existing library.” In this installation, a real-life space in one country is reconstructed in another, making the role of Deutsches Informationsbibliothek Pjöngjang both real (it did exist) and unreal (it is reconstructed for this installation). How those creating metadata should deal with the concept of the Deutsches Informationsbibliothek Pjöngjang is the key ethical issue under discussion.

The reconstruction element of the Deutsches Informationsbibliothek Pjöngjang is described within the documents being cataloged. Nevertheless, the fact that this is not actually the original Deutsches Informationsbibliothek Pjöngjang nor takes place in Pyongyang is not obvious from the title of the documents. Arguably this requires some careful reading to unpick, and there is a possibility that the title of the work could be misleading to users. Furthermore, though the exhibition’s location on the title pages makes it obvious that this did not take place in Pyongyang, it is not obvious that Deutsches Informationsbibliothek Pjöngjang really did exist. Therefore, if Snow’s (2015) multiple realities were employed here, then a combination of the title and subject of Deutsches Informationsbibliothek Pjöngjang (self-representation, ICP 2.3) with a note explaining the reconstruction element (accurate portrayal, ICP 2.4) would be a helpful solution and of maximum convenience to the user. So, in this example, the artistic unreal can be adequately dealt with using Snow’s (2015) conclusions for false memoirs.

3.3 Example 2: The Pyongyang Times by the Random Institute

The second example is entitled The Pyongyang Times. This is documentation from an exhibition that took place April 9-12, 2016, on the 23rd floor of the Yanggakdo International Hotel in Pjöngjang, North Korea (Random Institute 2016). The exhibition was entitled “All the lights we cannot see,” and featured the work of nine international artists (Random Institute 2016). The exhibition was arranged by the Random Institute in Zürich, who also published the catalog, and the exhibition was curated by Anna Hugo and Sandino Scheidegger (Random Institute 2016). There are two broad sets of aesthetic-ethical KO questions raised by this example: the manipulated aspect of the documentation and the issue of secrecy around the exhibition.

The printed documentation takes the form of a newspaper entitled The Pyongyang Times. There is an article about the art exhibition on page three while the rest of the issue deals with other news, politics, culture, and so on. However, the catalog is actually a manipulated version of this issue of The Pyongyang Times, an English-language newspaper produced in North Korea. In the most part, the catalog replicates the actual issue of the newspaper, including its masthead, numbering (No. 15, 2894) and date (April 15, 2016) and the majority of the articles. The main difference between the real Pyongyang Times (2016) and the unreal catalog (Random Institute 2016) is the insertion of the article about the exhibition, which apes the stylistic conventions of the host publication and replaces a story about pump factories by Sun (2016). Less explicit signs appear at the end of the publication that this is not the newspaper Pyongyang Times; there is an attribution to the Random Institute and a handwritten copy number. Ultimately, this example is different in essence from examples such as false memoirs, as the website for the catalog (Random Institute 2020) is transparent about this manipulation, and there are small indications on the catalog itself.

The manipulated nature of this documentation gives rise to a question about the most suitable title and also whether its manipulation should be explained within the catalog description. The exhibition documentation has a title of The Pyongyang Times on the documentation; yet, there is a question about whether this is an accurate portrayal of the
title, considering it is a manipulation of this newspaper. Applying Snow’s (2015) conclusion about multiple realities helping the convenience of readers, at first, seems useful for this scenario; this would see the title in the metadata as The Pyongyang Times with the addition of a note describing the manipulation and a link (with appropriate designator) to The Pyongyang Times (newspaper). However, creators’ intentions come into play; it is assumed that the creators want to give the illusion that there was an article about this exhibition in The Pyongyang Times (newspaper). So, notes and links that foretell the manipulation within this document also eliminate the reader’s opportunity to experience this unreal world, if the reader approaches the document through the library catalog. This sets up an interesting relationship between creator, cataloger and user.

Secrecy also presents ethical considerations for those making this resource discoverable. The Random Institute (2020) states that “the entire project, from the art on display to the story behind it, was committed to silence.” Random Institute (2020) also says that the only surviving documents are some installation views and a short mention on the CVs of artists who participated, who were then not allowed to say anything about the exhibition when they would inevitably be asked to elaborate. However, there is definitely an online presence for the exhibition, which appears to have been perpetrated by the authors; for instance, the Artcritical site (Random Institute 2020) and reviews (for example, Starnes 2016). The secondary source by the authors (Random Institute) juxtaposes the details of the exhibition with the possible unreal element of the movement, plus a further question about the unreal nature of the organization as part of the KO ethics question in these documents relates to the Jewish Renaissance Movement in Poland; there is a question about the realness of this movement, which then generates a question about how to represent this potential un-realness in the metadata associated with the posters. According to the manifesto (Bartana 2009a), the Jewish Renaissance Movement in Poland was an organization whose purpose was to organize a return to Poland of the Jewish population. However, it seems that in 2009 this movement existed only within the universe of Bartana’s artworks. According to the definition of a movement in the Oxford English Dictionary (“Movement, n., 8a” 2003), a key ingredient in a movement is that it must involve multiple people working towards the same thing; the Bartana example does not appear to have this multiple aspect.

The question of ethics arises when contemplating a note about the unreal nature of the organization as part of the metadata and then whether to potentially add more legitimacy to the organization by indexing it. Following ICP 2.3 (self-representation) would mean not including a note questioning the movement’s realness and that the movement should be indexed. Conversely, following ICP 2.4 (accurate portrayal) would mean including both a note about the possible unreal element of the movement, plus a further decision to be made about whether it should be indexed as a fictitious body or not at all. If multiple realities were adopted here (Snow 2015), there would be some extra complexities. First, this example is not just a question of whether
the real or unreal version should be represented in the metadata; instead, reality is blurred. The Jewish Renaissance Movement in Poland is real-ish, but our cataloging systems cannot (usually) account for this ambiguity. Second, the artist’s wishes are significant; following a combined approach also means changing the intended experience of the user and probably going against the wishes of the creator. Furthermore, unlike the previous examples, Jewish Renaissance Movement in Poland’s unreal(ish) status is not explicitly commented upon in the exhibition documentation and is not easy to ascertain through authoritative sources. In fact, its realness is quite a struggle to establish, to the extent that it is reliant on conjecture. This suggests that transparency about a deception can vary, even within artistic examples. It could be hypothesized that by not acknowledging the unreal element, the creator is more committed to this unreal experience and perhaps their wishes take on more, not less, importance when balancing up the cataloger’s competing considerations.

However, there is a twist to this particular example. Bartana continued working on these themes; the movement that was unreal-ish in 2009 later becomes real. Cembalet (2013) said the artist “brought her campaign into another reality when she turned the Jewish Renaissance Movement into a real entity.” Hence, there is now a question about whether the metadata should be altered post-facto for this work to represent what has since happened to the movement. This is particularly interesting when compared to the example of false memoirs. For example, Snow (2015) discusses the ethical responsibilities around updating the metadata when catalogers find that a work that describes itself as factual is later found to be partially or entirely fictional. However, the Bartana example presents a different situation. In the Bartana example, the reality changed with time; for false memoirs, it is the information known about the reality that has been updated. So, the Bartana examples emphasize the consequences of cataloging happening in a temporal frame and the importance to KO of understanding the relationship between time, reality and information about that reality.

3.5 Impact on specific metadata elements

The discussion about these three examples of art documentation has described the broad areas and questions raised. It is now worth briefly considering the specific elements that might be impacted by these questions. The specific types of information that emerged from the analysis of the three examples is listed in the first column of Table 1, and for each specific type of information, the example(s) it appeared in is also indicated. To give some useful context, the equivalent or near-equivalent RDA attribute(s) or relationship is given in the second column alongside its current guideline numbering (“RDA Toolkit” 2020). Additionally, the corresponding MARC21 field is also given in the third column (Library of Congress 2020b), and where the RDA element is only mapped to specific MARC21 subfields, the specific subfields are also specified using the convention of a dollar sign (Library of Congress 2020b). While most values in the MARC21 column refer to bibliographic metadata, one element is also concerned with authority metadata (Library of Congress 2020a).

We can see a number of important ideas from Table 1. It appears that the information from these three examples is centered upon quite a few different areas, including title, publication, notes, contributors and related works. However, if more works were added to future analysis, undoubtedly more elements would emerge, so this table is not at all exhaustive. The extremely small number of examples means we can make no assumptions about the frequency of the different elements; however, from this information, it is sensible to suggest that elements related to title, notes and related agents or works appear to be reasonably likely to be affected by unreal-ness in exhibition documentation. It is also interesting to note that all of these examples of elements would usually be user-facing rather than for internal use.

A further question emerges regarding classification and subject indexing. Exhibition documentation is treated differently in various global, general classification schemes, and art libraries are also known for their high incidence of adapting classification schemes (Ferrari 2000; Lee 2011). Ultimately, the impact of unreal elements on the classification of exhibition documentation would depend on the organization’s classification scheme and any local adaptations, although the above three examples are unlikely to be impacted by classification schemes, which chiefly categorize by solo artist, artistic medium or exhibition venue. However, format categorization might have an impact for The Pyongyang Times example, where the ethical considerations around manipulation, findability and secrecy—alongside other practical considerations to do with preservation and size—might influence the decision about whether to categorize this art documentation alongside other print newspapers. The three examples also show how considering unreal aspects might have an impact on subject indexing, including whether to include an unreal organization as a subject or not, and if so, whether to indicate the fictitious nature of the organization within the index term. So, while this discussion does not specify exactly which elements and aspects of knowledge organization will be affected by unreal aspects, it does illuminate some of the possible ways that unreal aspects of art documentation works might manifest themselves in an applied setting.
Examining these three examples of art documentation has teased out some important findings. Two key ideas emerge: 1) the importance of creativity and the role of the creators in making ethical KO decisions; and, 2) that unreality and reality are important to KO and are complex concepts. These are now built into three models that focus on artistic reality while also expanding our general understanding of dealing with deception and unreality in KO.

### 4.1 Model one: creators’ wishes within the ethics of knowledge organization

The analysis of the three examples highlighted a significant consideration: the position of the creator when dealing with unreal-ness in art documentation. However, before examining this in detail, it is useful to contemplate why the artists’ wishes should be accommodated. In other artforms, honoring the wishes of the creator is so inbuilt into the artistic process that it has historically been an assumption; for instance, in musical performance, the need to obey the composer’s intentions was an almost unquestioned assumption (Dipert 1980) until recent decades. While artworks mostly do not have the same dependence on intermediaries as the performing arts, some of these artistic examples do share an essence of this performative aspect and thus this assumption could be borrowed. Furthermore, there is also a residual moral imperative to respect the wishes of creators, which is in some aspects is also enshrined in intellectual property law.

The examples illuminated the role of the creator in the ethics of KO, and furthermore, how fulfilling the creator’s wishes could be in opposition to other ethical principles such as accuracy. For example, in The Pyongyang Times example, the needs of the creator would favor honoring the artistic experience for users, which means describing the document as though it were a real issue of the newspaper. Yet, doing this would arguably not be an accurate portrayal of the document and could lead to inconvenience for those seeking the document via its exhibition name or those actually wanting the Pyongyang Times (newspaper) (Here, the term “inconvenience” is used to express the situation when a reader looks for something on a catalog, and the search results are either longer than they would be or do not include the desired document, due to the decisions made by the cataloger around title). The idea of an obligation to creators is not present, or at least not explicit, in various discussions about ethical obligations in cataloging. As discussed in the literature review, Bair (2005) discusses groups to whom catalogers have obligations: society, employer, clients, colleagues and other professional organizations, the profession as a whole and individuals. The creators of the bibliographic work are not mentioned. Therefore, artistic unreal works uncover a potential additional ethical area and obligation.

### 4.2 Model two: unreal/unrealish organization as subject

The area of creators’ wishes and visions opens up a rich seam of relationships, complexities and understanding. First, there is clearly an important relationship between the creators’ wishes and the convenience of the readers. Honoring the creators’ wishes may mean possible inconvenience to the readers, especially if information is omitted. This sets up a vector of importance, and starts asking how metadata cre-
ators prioritize the competing needs of the various groups listed in Bair (2005) and others (Note that in the examples seen, the purpose of following the creators’ artistic vision is also to benefit the readers, as it is their experience of the documentation which is at stake). Second, it is not necessarily straightforward to determine the obligations to the creator as they are likely to be implicit. For example, the creators’ desires for The Pyongyang Times are implied from looking at the spaces between the document (looks like a newspaper) and what it actually is (a manipulated document), combined with secondary sources and general broad ideas about art. The proposal that the creators want the readers to think (at least in the first instance) that this is a real copy of the newspaper The Pyongyang Times is an assumption, so all the subsequent cataloguing implications are based on this assumption. Third, The Pyongyang Times example highlights that there is a difference between the document and the creator in terms of ethical considerations. In other words, just following the document alone will not necessarily also take into account the document creators’ wishes.

So, a fourth ethical obligation could be added to the three identified by Snow (2015): wishes of the creator (Note that this is given in the singular to match the language used in the ICP obligations, which talks about user rather than users). Snow’s (2015) framework and research for deceptive works suggested that the best way to achieve the most convenience to the user is a combination of self-description and accurate portrayal, which could be written as “self-representation (ICP 2.3) + accuracy (ICP 2.4) = best convenience to the user (ICP 2.1).” Adding the creators’ wishes extends this as follows: “self-representation (ICP 2.3) + accuracy (ICP 2.4) = best convenience to the user (ICP 2.1) + creators’ wishes (new).” In other words, the required result is to inconvenience the reader as little as possible while also following the projected wishes of the creators; this is achieved through some combination of self-description and accurate portrayal. This is visualized in Figure 2, which provides a model of the competing ethical demands and shows how these interact with the bibliographic description and organization. This extends a similar visualization seen in Figure 1.

Figure 2. Model one: creators’ wishes within the ethics of KO.
to include the extra balancing needed to incorporate the obligation to creators. The model also shows how the principles on the left-hand side pertain to the document while those on the right pertain to different sets of agents (using the language of IFLA-LRM here). It could be argued that the authors of false memoirs actually have similar wishes to the creators of the three analyzed artistic examples. The difference between the situations is one of motive and this difference suggests that there are other factors at play in terms of unreality. These will now be explored further.

4.2 Model two: concentric circles of unreality

The exploration of the three examples highlights that the unreal element is not uniform in art documentation. For example, The Pyongyang Times had unreal aspects in the format of the exhibition documentation, which borrowed an existing publication; the Heide example had its unreal element within the art itself, where the artwork being discussed in the exhibition documents was an amalgamation of the real and unreal; the Bartana example’s unreal-ness is in the subject of the exhibition. It is possible that ethical decisions about the KO of such documents might differ depending on where the unreality sits; potential misleading in the title of a document (for instance, The Pyongyang Times) might lead to a bigger imperative to explain the unreal-ness to the user than an unreal element in the artworks (for instance, Heide) or the subject of the artwork (for example, Bartana). This is partly to do with expectation; artworks are expected to have creative interpretations of reality—though, not to the extent of say a novel—but this is less expected in the art’s documentation. Furthermore, the difference between the three examples is the distance between the unreal aspect in relation to the act of metadata creation and indexing. This could be considered a “metadata librarian’s gaze.” This links to the question of what exactly is being cataloged in exhibition documentation, which sees contemporary bibliographic cataloging practices focus on art documentation rather than the artwork or exhibition. Three types of unreality are visualized in Figure 3. This delineates unreal elements of art documentation in comparison to unreal elements within the artworks, showing how unreal elements in the artwork are more removed from

Figure 3. Model two: concentric circles of unreality.
the bibliographic catalog records than those that exist as part of the documentation of those artworks. Unreal subjects of artworks are also represented in Figure 3, as part of the artwork but a further step away. Understanding the different placings of the unreality helps us to understand why different unreal elements may be treated differently by those creating metadata.

4.3 Model three: unreal-ness as a type of information

The artistic unreal examples demonstrate that there is an important factor for making decisions about a document’s metadata: motive for unreal-ness. In other words, why is this unreal or deceptive element in the document or artwork? In the three examples of art documentation, we assume that the creators’ reasons for the unreal aspects are related to the experiential elements of the artworks and/or exhibition documentation. This is in contrast with, say, false memoirs, where the unreal element may be to represent the author in a better light, to increase sales or that the author may not be able to distinguish fact from fiction, and so on (This leads to a question that is outside the remit of this paper about ability to distinguish fact from fiction, and so on (This leads to a question that is outside the remit of this paper about whether it is possible that one person’s false memoir is another person’s artistic deception).

So, unreal-ness as a category of information is visualized in Figure 4. Motive appears as one of the constituents feeding into unreal-ness. Some example classes within the category of motivation are given, such as deception for artistic purpose. As discussed, different motives for the unreal elements might lead to different practices by metadata creators. The importance and closeness of the creator to the idea of motive is illustrated. This follows on from discussions earlier in this section about how artistic unreal-ness highlights the creators’ role in the ethics of KO. Another factor feeds into unreal-ness: transparency. The three artistic examples highlight there is a scale of the ease in locating the unreality. For example, the Heide example had explicit information about the unreal element on the actual exhibition document, whereas similar information was very difficult to find for the Bartana example. All three art documentation examples are in contrast to other types of unreal works such as false memoirs, where the transparency value would be zero. Transparency would be useful information for metadata creators when deciding whether the unreal-ness should be included in the description at all, and if so, could inform how the metadata for these elements is treated. Figure 4 illuminates possible connections between the motive for deception and the transparency. Motives such as improving sales or reputation are likely to correlate to low transparency; conversely, artistic purposes are visualized as linking to medium and high transparency (acknowledging that the Bartana example, which does not contain explicit statements about unreal-ness, is a counterexample to this).

The resulting unreal-ness of the document manifests itself in different aspects of the resource, which are shown as the “element which is unreal” category in Figure 4. Brubaker (2018) distinguishes between deceptions in authorial information and deceptions within the documents, so these types of information are included as classes within the “element which is unreal” category. Added to these are the classes of artwork and subject of artwork (as discussed above and seen in Figure 3). The inclusion of subject here means that, potentially, subject classification work is added to the areas of practice that are impacted by the findings in this article. A further class for exhibition highlights the divisions between an exhibition and its documentation. Again, contemplating these aspects of the resource in relation to a non-artistic example is helpful; for instance, while artwork or exhibition information would not be useful for works such as false memoirs, subject and authorial information are especially pertinent when creating metadata for false memoirs. Ultimately, Figure 4 threads together much of the previous discussion about unreal-ness, and illustrates how artistic examples show the breadth and nuance of unreal-ness and its ethical implications.

5.0 Conclusion

Art documentation provides specific challenges to those attempting to describe and index them in an ethical fashion. This article has used three examples of documentation relating to exhibitions to highlight where fault lines can occur. These have illuminated differences to other types of deceptive works and presented new dimensions to the ethical issues of KO. The three examples show how for some art documentation, the documents are part of the experience of the artworks; in turn, the metadata creator then has to ponder whether to include metadata about unreal elements if that might have an impact on the user’s experience of those documents. So, the ethical needs of the documents’ creators become important. Therefore, the artistic examples and model one contribute to the development of discourse surrounding applied KO ethics: they add a novel ethical need, they introduce a novel set of people to add to the list of catalogers’ obligations and they add an extra layer to the complications of balancing different ethical principles. The examples also illuminate a hitherto unknown aspect of dealing with deceptive works: the richness of the idea of reality. Model two visualized the position of the unreal element in relation to the catalog record and introduced the idea that different cataloging and indexing decisions might be made depending on the distance between the unreal element and the metadata. This is a novel aspect of considering deception and unreality in KO ethics. Model three posited that unreal-ness is actually a whole type of information, belonging perhaps to meta-metadata, and that this category of in-
formation is critical for cataloging and indexing any item. Furthermore, though these models germinated in the artistic unreal examples, they are potentially applicable to any document that has an element of the unreal. Therefore, considering the ethics of KO for art documentation expands our knowledge and understanding of applied ethics more generally.

There are a number of wider uses and implications to these findings. The importance of considering art-related works in research about ethics is clear. Therefore, it would be valuable for future work on codes of ethics to specifically consider art documentation such as the examples presented. In a similar vein, this work is also important for those developing general guidelines about art cataloging. For instance, this ethical lens would be a valuable addition to future editions of manuals that discuss cataloging exhibition documentation, such as guides by the British Library (2020) and ARLIS UK & Ireland (2000). The discussions about unreality are potentially useful when considering the theory of the practice of metadata creation, as it brings in unreality as a new, complex aspect of information and asks questions about what, exactly, is being described and indexed. Finally, the ideas presented in this article are also valuable when thinking about more general issues to do with the findability and display of metadata to users. For instance, while theoretically there is no way to simultaneously include a piece of information (it is not the original newspaper) and to not include it (it is the original newspaper), there is a solution from a systems perspective. The resource discovery layer or catalog could be designed to blank out specific information related to unrealties, which the user than chooses to uncover; this is akin to “spoiler sections” found in some resources about film and television. So, these art examples are also insightful about potential interactions between KO decisions and user experiences of catalogs, and this could have implications for future systems design.

Future research in this area could explore other creative works. It would be valuable to consider art documentation in comparison to other artistic areas such as novels, films and music. Another useful extension would be to consider the ethics of creating metadata for pastiche and parody and whether similar issues and structures emerge to those discussed here. It would also be fascinating to link these findings to research in machine learning in bibliographic de-

Figure 4. Model three: unreal-ness as a type of information.
scription and knowledge organization. The Bartana example showed how difficult it was for a trained human cataloger to establish whether an aspect was real or not. What does this mean for description and indexing produced by machine learning? Ultimately, considering the KO ethics of art documentation not only extends our thinking about how to deal with deception and unreality from a KO perspective but also reframes the broader relationship between metadata creator, artistic creator and library user.

References


OED Online, s.v. “Deceive, v., 2a.”
OED Online, s.v. “Fiction, n., 4a.”
OED Online, s.v. “Movement, n., 8a”
OED Online, s.v. “Unreal, adj. (and int.) and n., 1.”
Pyongyang Times. 2016. 9 April (No. 15, 2894).
The Abandonment of the Assignment of Subject Headings and Classification Codes in University Libraries Due to the Massive Emergence of Electronic Books

Isidoro Gil-Leiva*, Mariângela Spotti Lopes Fujita**, Pedro Díaz Ortuño*** and Daniela Majorie dos Reis****

Universidad de Murcia, Facultad de Comunicación y Documentación, 30100, Espinardo, España,
* <isgil@um.es>, *** <diazor>
Universidade Estadual Paulista Júlio de Mesquita Filho, Marília, SP, Brasil,
**<mariangela.fujita@unesp.br>, **** <danielamajorie@yahoo.com.br>


Abstract: The massive and unstoppable emergence of electronic books in libraries has altered their organization. This disruptive technology has led to structural changes. Currently, an e-book exists only if its metadata exists. The objective of this article is to analyse the impact that the massive incorporation of electronic books in university library systems is having in the processes of assignment of subject headings and classification codes. We carried out a survey of more than six hundred libraries, which means almost all the university libraries in Portugal, Spain, England, United States, Brazil, Sweden, Norway, Finland and Australia. From the results obtained, it is deduced that: 1) librarians expect e-books to be provided with descriptive metadata related to the subject headings and classification codes; 2) the bibliographic records provided by publishers/providers seem to be improvable; 3) the quality of the metadata provided by the providers does not seem to be taken into account when selecting publishers for the purchase; 4) the discovery tools
are also clearly improvable; 5) it seems that there is no “frustration” or “stress” among librarians about the changes produced in relation to technical processes; and, 6) it does not seem that we are facing a paradigm shift motivated by these issues.

1.0 Introduction.

Indexing (understood as a conceptual representation intended to satisfy information needs) implies an indissoluble union between the processes of indexing, retrieving and using information. This connection between indexing and retrieval has been occurring since the moment a conceptual representation was used for information storage purposes. Three events have been key in shaping this union. The first occurred in Mesopotamia with the use of clay tablets, but it was consolidated with the appearance of medieval monastic libraries and later, between the sixteenth and eighteenth centuries, with the incipient libraries that were first the product of humanist ideas and later of the scientific revolution, when more complex procedural and organizational techniques began to be applied.

A third phase took place at the end of the nineteenth century with the introduction of tools such as A Classification and Subject Index for Cataloguing and Arranging the Books and Pamphlets of a Library by Melvil Dewey and the Rules for a printed dictionary catalogue by Charles Ammi Cutter, which formed the theoretical and practical foundations of subject assignment and classification in libraries by combining information storage and retrieval.

Librarians created mechanisms to easily locate books using both classifications and catalogues that acted as access points governed by author names, which facilitated knowledge of all the bibliographies of an author available in the library, the titles catalogue, which facilitated the location of a book or made it possible to discover unknown titles and, finally, the subject catalogue which ensured quick and efficient access to all the books in the library on a given subject. This was the discovery system for a long time in libraries. In fact, the traditional author and subject catalogues have conceptually survived until today since, from any bibliographic record that we are visualizing on a screen, with a simple click we can obtain a list of all the works of an author or a list with all the resources on a certain subject. Similarly, classifications following hierarchical principles have been beacons in the conceptual organization of works by fields of knowledge and their physical location on the shelves (Dewey Decimal Classification, Library of Congress Classification, Universal Decimal Classification, Colon classification or Bliss bibliographic classification).

In the early 2000s, e-books began to arrive in libraries intensively and unstoppably. A few decades earlier, Andries Van Dam, coined the term “electronic book” while working on the first hypertext system back in 1967, while Alan Kay, introduced the Dynabook, an e-book he imagined as an interactive portable personal computer with a flat-screen and wireless communications (Ardito 2000). Likewise, for Connaway and Wicht (2007, 1) there are several milestones and initiatives that are part of the history of e-books: the ideas of Michael Hart in 1971 that later became Project Gutenberg, the proposals of some publishers and vendors who in the late 1990s started hosting and selling e-books, the emergence of e-book providers such as E-Book Library (EBL) or MyiLibrary, which started offering flexible pricing and innovative access models, or the announcement in late 2004 by Google of its Google Print Library Project (later renamed Google Books Library Project) in cooperation with several libraries.

At present, it seems that some traditional library approaches, services and processes are becoming increasingly irrelevant (Stachokas 2014, 33). On the other hand, a growing number of librarians all over the world are beginning to recognize that libraries are not properly organized for the work that actually has to be done in the twenty-first century (Barnes, as cited in Stachokas 2009, 207). Parallel to this (Breeding 2017, 16), “in most every respect, digital and electronic collections have proven to be more complex than their analog predecessors.”

Furthermore, Stachokas (2019, 41) notes, in referring to the transition from the traditional library to the hybrid library, that a different approach has been required in terms of acquisitions, cataloguing, access, management and evaluation. In this sense, it is worth highlighting here what Goedeken and Lawson (2015, 2016) wrote: “in the last centuries, libraries developed sophisticated bibliographic structures to accommodate the printed book and its acquisition, description and classification. In the period of two decades, however, this well-established arrangement has been shaken by the disruptive technology that e-books have implied.”

It has already been pointed out that one of the elements that have contributed to the extension of e-books in libraries was the appearance of innovative acquisition/subscription systems, a factor that has led to massive incorporation of e-books into university libraries initially and later in libraries in general. This has meant that, in a short time, tens of thou-
sands of e-books have been incorporated into library systems.

To illustrate this, we bring here an example that may be representative of what is happening or will happen soon in most of the world’s libraries. This is statistical data from Rebiun (Spanish university libraries network). Some of the decreases that may be seen in some articles are likely due to the non-contribution of data by one or more libraries.

According to Table 1, while printed books (hereinafter referred to as p-books) increase in number by approximately 2%, e-books do so by 6-10% each year, although the number of librarians appears to remain completely stagnant in the period.

From this simple statistical data, it is intuited that perhaps librarians have been facing an increase in work and over-head in the management of p-books and e-books. Martin and Mundle (2010, 227) state that “cataloging individual books may be impossible when large packages are purchased. Increasingly, libraries are relying on outside sources for their e-book catalog records, which may come from vendors or third-party record services and are frequently included in the price of a subscription.” To this new dynamic that has been installed in libraries, we must add the shortage of catalogers. In 2012, Sapon-White (2012, 46), Head of Cataloguing and Metadata Services of Oregon State University, wrote that the university had 19,000 students and 1,800 faculty members and that libraries annually acquired some 15,000 monographs and about 5,000 government documents. However, libraries had only a paraprofessional dedicated to downloading cataloging records and a head of cataloging as the sole “professional cataloger in the unit, with time split between administering the unit and providing original and complex copy cataloging.” Although this is a very concrete example, it does not seem that a “paraprofessional” and a cataloguer was the optimal team to carry out the great amount of work that can be intuited. Therefore, these new work dynamics that have brought the incorporation of e-books to the libraries and precarious library staff lead us to consider whether today, university libraries have the capacity to just check the metadata that accompanies e-books or whether librarians have the time to correct and improve the metadata of tens of thousands of electronic bibliographic records provided by vendors.

Perhaps, this work overload of librarians could be affecting two central and well-established processes in libraries that had been determining both the physical organization of books through classification and, the use of bibliographic resources by users through retrieval by means of subject headings. Until the emergence of e-books these essential technical processes had been carried out with a widespread rigor in libraries. With the gradual arrival of e-books but painstakingly with their massive incorporation into libraries and the overload of work that librarians seem to carry, a question arises: how the execution of these key processes has been or is being transferred to the e-books. Bearing this in mind, we consider that research is needed to know specifically how the technical processes of assignment of headings and classification codes to e-books are being carried out and to find out what librarians think about issues related to these aspects. We consider that this research is necessary not only because of the data and information it can provide but also because we are facing processes that directly affect the location and use of resources by users. These technical processes directly affect users, because if the e-books arrive at libraries with deficient, incomplete or erroneous indexing and classification metadata and librarians do not have the capacity to check (due to lack of time and staff)—and if we add to this the reception of possible deficiencies or errors in the title or author metadata—bibliographic records become invisible to discovery systems, and, therefore, what does not exist cannot be located first or used later. Thus, we consider it relevant to have a study on these issues in university libraries, because they were the first and the ones that, with most intensity, have been massively incorporating e-books.

This work is intended to continue and expand a study initiated in 2016 and published in Proceedings of the Fifteenth International ISKO Conference (Gil-Leiva et al 2018). In that work we presented the results of the analysis of a web questionnaire sent in December 2017 to libraries in Portugal, Spain, United Kingdom, United States and Brazil. In the work we present now, as we will explain below, the questionnaire has been sent to new university libraries in Sweden, Norway, Finland and Australia, so a larger set of data has been achieved and a broader presentation, contextualization, analysis and discussion of all the data has been ob-

<table>
<thead>
<tr>
<th></th>
<th>2013</th>
<th>2014</th>
<th>2015</th>
<th>2016</th>
<th>2017</th>
<th>2018</th>
</tr>
</thead>
<tbody>
<tr>
<td>E-books</td>
<td>8,819,673</td>
<td>9,452,026 (+6.69%)</td>
<td>10,767,685 (+12.21%)</td>
<td>11,955,167 (+9.93%)</td>
<td>12,699,922 (+6.29%)</td>
<td>14,116,385 (+10.30%)</td>
</tr>
<tr>
<td>Printed monographs</td>
<td>27,865,360</td>
<td>26,994,464 (+1,90)</td>
<td>28,319,147 (+1,35)</td>
<td>28,869,028 (+1,90)</td>
<td>29,266,657 (+1,35)</td>
<td>29,369,554 +0.35</td>
</tr>
<tr>
<td>Librarians</td>
<td>2400</td>
<td>2416 (498)</td>
<td>2391</td>
<td>2408</td>
<td>2416</td>
<td>2418</td>
</tr>
<tr>
<td>Courses given</td>
<td>445</td>
<td>452</td>
<td>718</td>
<td>644</td>
<td>498</td>
<td>597</td>
</tr>
</tbody>
</table>

Table 1. Statistical data of the Spanish university library network. Source: https://rebiun.um.es/rebiun/admin/ManageIndicatorsPage
2.0 Literature review

Over the past two decades, both LIS researchers and librarians have worked and published extensively on numerous aspects of the library and e-book ecosystem, such as platforms, demand services and e-book sharing networks, challenges faced by libraries in this format, acquisition models, discovery tools, loans, e-books versus p-books, support in assisting in selection and purchase, uses, challenges and changes in cataloguing or quality of bibliographic records provided by suppliers, among others. Thus, without pretending to carry out an exhaustive bibliographic review since it would exceed the scope of this work, we will limit ourselves to some of these aspects that are more related, directly or indirectly, to our research, and which we group in seven sections.

2.1 Challenges and changes in e-book cataloguing

Nowadays cataloguers find themselves managing and manipulating large sets of catalogue records, but batch processing brings about a new set of challenges to expedite the process of getting vendor records into the catalogue. In this sense, work has been done to describe experiences in libraries about management and cataloguing practices for e-books such as batch-loading, workflow, quality control (Martin and Mundle 2010 (a); Zhao and Zhao 2010; Chen et al. 2016; Castro, et al. 2019).

2.2 Quality and quantity of metadata provided by suppliers

This issue, which is related to our research, has also been addressed in several publications (Wu and Mitchell 2010; Zhao and Zhao 2010; Park and Tosaka 2010; Ravit and Dana 2015; Frederick 2016; Wiersma and Towstidi 2017 or Yuan et al. 2018). Frederick (2016, 68) noted that e-books require a robust and accurate bibliographic description, although we believe that this robust and accurate bibliographic description is not occurring in library environments. In this sense, Sánchez et al. (2006, 55), when checking NeLibrary records, detected errors in titles, formatting problems, and the absence of certain MARC fields. Wu and Mitchell (2010, 171) according to their experience in cataloguing e-books at the University of Houston libraries, reported that a significant number of e-book bibliographic records provided by suppliers contain syntax and content errors. Elsewhere, Zhao and Zhao (2010, 98) analyzed MARC records from various suppliers and detected the existence of general, unauthorized, and spellbound subject headings on the SpringerLink platform. Other significant examples were shown by Wiersma and Toivstadi (2017, 628) who analyzed the descriptive content (keywords or subject headings) of certain e-books on various platforms and found that the differences are so great as to make us think that we are looking at different books. And finally, according to Breedt and Walter (2012, 5) and Wiersma and Toivstadi (2017, 624), e-book publishers seem to be aware that they are providing bibliographic records that can be improved, and in the same way, they know that these data gaps and deficiencies affect their prestige and sales. Work has also been done to identify potential problems and strategies for data cleansing and editing before making records accessible to users (Sanchez et al. 2006; Mi and Pollock 2019).

2.3 Challenges for libraries concerning e-books

Work has also been done on the integration of e-books in libraries (Connaway and Wicht 2007); the description and development of digital services platforms to support workflows and shared data models (Buczynski 2010); changes in the acquisition and circulation of e-books (Rodrigues and Godoy Viera 2018); applications for downloading e-books by users (Wu et al. 2015); development of APP to enable users to read e-books from multiple vendors through a single interface (Beswick et al. 2017); interlibrary loan of e-books (Sewell and Link 2017); analysis of the roles, workflows of e-book life cycle and recommendations regarding the applications of various types of metadata (Lagace 2018); emphasis on shared values, building an ecosystem of interoperable platforms and tools and challenges as preservation, discovery and accessibility (Watkinson 2018); promoting awareness of e-book collections and enabling browsing and serendipitous discovery (Tingle and Teeter 2018); or the emergence of new technicians such as the electronic resource librarian (ERL) (Stachokes 2009 and 2019).

2.4 Acquisition models

The different options available for the acquisition of e-books have been described (Costello 2017; Serra and Segundo 2017); some recent experiences of PDA/DDA acquisition are: (Herrera 2012; Goedeken and Lawson 2015; Bennett 2016; Schroeder and Boughan 2018; Brown and Currie 2019); and demand-driven acquisition program of print books (England and Anderson 2019).
igate all available content and user-guided navigation to encourage serendipitous discovery (Tingle and Teeter 2018), user experiences in browsing collections using discovery tools (Bardeen et al. 2017) or the enumeration of requirements for navigation interfaces (McKay et al. 2018).

2.6 E-books versus p-books

This topic has also been extensively covered in several studies, so only a few examples are brought here. Thus, the use of p-books and e-books for identical titles in a collection has been compared (Goodwin 2014; Lewellen et al. 2016); comparison of the influence of e-books and p-books on reading comprehension, eye strain and perception (Jeong 2012); reading preferences of college students over e-books or p-books (Aharony and Bar-Ilan 2018); how format (e-book/p-book) affects the use of e-books (Miller 2014; Fry 2018; Yuan et al. 2018); pricing of e-books versus p-books (Rao et al. 2018); or the relationship between the technology acceptance model and the preference for e-books (Smith et al. 2019).

2.7 Uses of e-books

Special attention has been dedicated to this topic and countless studies on perceptions, behaviour patterns and use of e-books have been published. Some examples are Noorhidawati and Gibb (2008), Lam et al. (2009), Croft and Davis (2010), Smyth and Carlin (2012), Romero-Otero et al. (2013), Gray and Howard (2017), Blummer and Kenton (2018), Casselden and Pears (2019); guidelines for librarians on the use of e-books in reference service and collection development (Rokusek and Cooke 2019; Proctor 2019); relationship between licensing and interlibrary loan (Fredderiksen et al. 2011); relationship between academic e-book user practices and selection, browsing and reading (McKay 2011); university student perceptions of e-books versus printed books (Rojeski 2012; Rafiq and Warraich 2016); impact of the book’s format on loans (McKay et al. 2015); use of mobile phone for reading e-books (Wang et al. 2018); the relationship between the technology acceptance model and preference for e-books (Aharony 2014; Smith et al. 2019); the effect of instruction on e-textbook use (Ragan et al. 2019). Elsewhere, work has also been done on the relationship between attitudes and use of e-books and to contextualize this in relation to sociodemographic background (Bergström 2018) or why students prefer both e-books and p-books (Noorhidawati and Gibb 2008; Croft and Davis 2010; Schonfeld 2013; Academic Student E-book Experience Survey 2018; Yuan et al. 2018; McKiel 2016 or Aharony and Bar-Ilan 2018).

3.0 Methodology

This research is based on the following sources and methods. First, a bibliographic review was carried out based on information searches in the Web of Science, Scopus and internet databases. Likewise, we prepared a web questionnaire on an internal platform of the University of Murcia with twenty-five matrix questions and closed questions. The content of the questions centered on the technical processes in e-books (assignment of subject headings and classification codes), discovery tools, collection stability, paradigm shift and frustration or stress among librarians and automation of the assignment of subjects and classification codes (Annex 1). This same articulation of content has also been used to present the results.

Before the mass mailing of the survey, we performed a pretest. We met with two librarians to discuss different aspects such as understanding of the questions, their suitability or their sequencing. Once the survey was prepared, we had to decide where and to whom to send it. Our intention was to gather as much information as possible and, if possible, from different geographical areas. Hence, the decision to translate the survey into three languages: English, Portuguese and Spanish. In this way, we could reach different countries and socio-cultural contexts. The English-language survey was sent to university libraries in the United Kingdom, Sweden, Norway, Finland, the United States and Australia. The Portuguese survey was sent to libraries in Portugal and Brazil, and finally, a Spanish survey went to Spanish university libraries. Furthermore, with the survey in Spanish, we could extend the research to countries in Central and South America.

As to whom to target with the survey, we decided that the best recipients would be the directors of the university libraries, or those responsible for the technical processes. So, we focused on getting the emails from these recipients. For each of the selected countries we searched the web for national directories of university libraries, and once located, we visited each of the university pages until we obtained the email of the director, the person responsible for technical processes or in their absence a cataloguing librarian. In this way, we tried to guarantee that the survey would be completed by the most appropriate people (in the presentation of the survey we expressly indicated that if there were a better suited person to complete it, the email would be redirected) and that it would reach all the university libraries in each selected country.

Once the surveys were available in web format in all three languages and with the e-addresses, more than six hundred surveys were sent to university libraries in Portugal, Spain, the United States and Brazil in December 2017. After three weeks we had eighty questionnaires completed. And at the end of November 2018, a second submission was made with
the same questionnaire to one hundred and forty-two libraries, almost all university libraries in Sweden, Norway, Finland and Australia. After eight weeks we had obtained twenty-seven completed questionnaires. A total of seven hundred and fifty-three surveys were sent.

4.0 Results

As we have pointed out, we sent seven hundred and fifty-three surveys, and one hundred and seven responses were received, resulting in a response rate of 14.2%. The highest response rate was obtained from Spain with 43.2% of the surveys answered, followed by Sweden with 30.5%, while the lowest rates were obtained from Great Britain (4.8%) and the United States (7.05%). Although it may seem a relatively low response rate, we consider it enough for a first approach to this issue, taking into account also that one third of the questions contained the possibility of inserting observations and comments by the librarians. In this way, we obtained more than three hundred comments that provided us with very valuable data and information. Perhaps the low response rate is related to the dates of submission. The surveys were sent between the months of November and January, months possibly overloaded with work at the end and beginning of the year to which we would have to add holiday periods. Below we offer the results.

4.1 Technical processes in e-books: assignment of subject headings and classification codes

4.1.1 Assignment of subject headings

The assignment of subjects to e-books does not seem to be a widespread practice in university libraries according to the data in Figure 1.

Librarians informed us that they do not assign headings because of the instability of e-book collections, the existence of small staff or because some libraries do not yet have e-books. 31.8% of the librarians assign headings to only some of the e-books. As shown in Figure 1, there is some uniformity in the responses, however, the data from Spain and Brazil are noteworthy, because 37.5% and 54.1% respectively have indicated that “yes” they assign subject headings to e-books while the rest of the countries surveyed are mostly between “no” and “some.” 24.3% of the librarians stated that on some occasion, after assigning subjects or classification codes and incorporating them into the e-book platform of the library, the publisher or supplier had eliminated or modified those records in overlapping processes in the common work areas. Elsewhere, they have indicated that due to the management policies of the libraries, and on occasions, to avoid being affected by this situation, they do not usually make changes or additions to the records that could be automatically modified; they do not assign subject headings or classification codes to the e-books if these are not provided by the suppliers, or they assign the subjects or classification codes only to those e-books that they are sure cannot be eliminated or modified by third parties. Finally, it is worth mentioning that 52.3% indicate that they do not know if, on any occasion, records have been modified by the publishers/suppliers.

4.1.1.1 Assignment of subject headings to p-books versus e-books

We also asked whether the process of assigning subject headings is the same as for printed books, and 85.7% of the librarians indicated that it is the same process, as many e-books also have the printed version; therefore, they found no reason to do it any differently. However, 14.3% indicated that

![Figure 1. Assignment of subject headings to e-books.](image-url)
it is not the same, because more generic subjects are assigned, global changes are made from the headings provided by the supplier or they are assigned from the title.

4.1.1.1.2 Process of assignment of subject headings

We also wanted to know the details of the process of the assignment of the headings. We asked whether they assign more or less the same number of headings for e-books as for p-books and whether there have been changes in the technical processes regarding indexing and classification. We also asked if they had been included in the library’s procedure manuals. Figure 2 provides the answers.

The high number of e-books subscribed/purchased by libraries makes an adequate allocation of subject headings unfeasible: this is the opinion of 54.36%, and almost 15% express doubt. On the other hand, 50.4% of respondents do not seem to agree with the assignment of subject headings in libraries, and almost 25% could be said to be undecided (Figure 3).

Similarly, 61.6% of librarians say that e-book subject headings should be provided by publishers while 19% is undecided (Figure 4).

Figure 2. Number of subject headings assigned to e-books and printed books and changes made to library procedure manuals.

Figure 3. Library should provide the subject headings.
In fact, 58% of libraries do not assign subject headings to e-books, because they are provided by publishers/providers. Likewise, 42% state that the subjects provided are sufficient and not worth modifying while 31% are undecided and 28% do not share the idea that they are sufficient. Also, 40% consider that they are adequate while 37% are unclear and 23% say that they are not adequate.

Regarding the statements on whether the metadata provided by the publishers/providers on subject headings are "adequate" and "sufficient and not worth modifying," almost all librarians seem to agree, and so they opted for options four to five on the scale, except for the Spanish librarians since 30% did align themselves with these majority options by marking four to five in relation to them being "adequate," but another 30% marked the opposite (disagree, options one to two). Similarly, in relation to whether "the subject headings are sufficient and not worth modifying," 41% of Spanish librarians are against this statement compared to 27.6% who seem to agree.

### 4.1.1.3 Provider selection

Finally, we were also interested in knowing if the quality of the metadata provided by the providers is considered by the librarians when selecting one or another e-books provider. 53.93% consider that the quality of the subject headings and classification codes provided by the publishers/providers is not considered when choosing e-book packages for purchase or subscription while 35.3% consider that it is.

### 4.1.2 Assignment of classification codes

Regarding the assignment of classification codes, 59.4% of the librarians considered that the high number of e-books purchased/subscribed in the library made an adequate classification number unfeasible while almost 27% declared they were against it. On the other hand, according to Figure 5, libraries assign classification codes purchased/subscribed in only 22% of cases to maintain the coherence of the catalogue since they are assigned to printed books.

Sometimes librarians assign classification symbols to e-books to represent those purchased individually and in property, depending on the quality of the data provided by the suppliers, or to e-books that are incorporated into the library catalogue and form part of the recommended bibliography in the teaching guides.

The participants were also asked to mark on a scale from one to five (1=Disagreement; 5=Agreement) whether the numbers codes (CN) provided by the suppliers are adequate and sufficient (Figure 6). From the answers given, there does not seem to be a clear consensus, although they are inclined to note that they are adequate and sufficient.

We also asked if they used the same procedure for assigning classification codes to e-books as for printed books and 81.1% indicated that it was the same. Several libraries stated that they assign classification codes either by making global changes to e-book packages based on correspondence with the general subject or by using tools to auto-generate the LCC. Finally, more than half of the librarians (56%) stated...
that the classification numbers should be provided by the e-book publishers/suppliers while only 22% considered that the librarians should execute this process.

4.1.2 Discovery tools

80% of the librarians consider that the assignment of subject headings to e-books is useful for the users during the handling of the discovery systems while 9.71% indicated that it is not and 11.65% did not know.

75.5% of university libraries have a single discovery tool for printed books, e-journals and e-books. The libraries of the Anglo-Saxon group (United Kingdom, United States and Australia) and of the Nordic group (Sweden, Norway and Finland) have uniformly answered that they have a single discovery tool. However, in the libraries of the Ibero-
American group, a certain divergence can be observed, since, in general, in the Spanish libraries there is a single discovery tool while in Brazil and Portugal half of the librarians indicate that there is no single discovery tool.

As for whether users are being offered easy and interactive options for discovering and searching for information, 48.1% said yes while 21.7% said no and 30.2% said they were not clear about this.

4.1.3 Collection stability, paradigm shift and frustration or stress among librarians

Sixty-six percent of those surveyed said that e-books are part of the university library just like p-books, although almost 30% said that they are not. On the other hand, there seems to be no agreement on the stability of e-book collections and their impact on catalogues according to Figure 7.

On the question of “whether e-book collections are unstable and make it difficult to keep the library catalogue up to date,” almost all countries scored mostly three and four, except Sweden where 50% scored one and three. To the question of whether there is a certain “frustration” or “stress” among the librarians involved in the tasks of subject assignment and classification of e-books, we have found divided opinions, as can be seen in Figure 8, although it seems that there is no “frustration” or “stress” about this issue.

On this question, almost all countries have opted for an intermediate position by marking option three, followed by option one. Only Brazil has its answers almost uniformly distributed throughout the scale from one to five. Annex 2
provides some of the reasons given by librarians who have expressed that they do feel some frustration or stress.

When asked if they consider that such significant changes are being produced in relation to the technical processes executed in the e-books (assignment of subject headings or classification numbers) such that we can speak of a change of paradigm or a change of model, there does not seem to be a clear agreement, as can be seen in Figure 9.

### 4.1.4 Automation of the assignment of subjects and classification codes

Almost half of the librarians indicated that it is appropriate to introduce tools for the automatic assignment of subject headings, while for the automation of classification code assignment the percentage is a little lower (41.58%). And in both cases, there is a remarkable 23% of indecision (Figure 10).

---

**Figure 9.** Paradigm shift due to changes in technical processes.

**Figure 10.** Need or need not to automate the assignment of subject headings and classification codes in libraries.
In general, all countries have positioned themselves between values three, four or five, so, regarding the automation of these processes, despite the noteworthy responses of Sweden and Brazil, where a clear tendency to disagree with the automation of these two processes is observed, since a significant part of their librarians opted for options one and two.

5.0 Discussion

P-books have traditionally formed stable collections in libraries. However, with the incorporation and management of e-book packages this idea of stability is changing, because some e-books or e-book collections seem to come in and out of libraries for different reasons. One-third of the librarians surveyed believe that e-books are not part of the library like printed books. This real or perceived instability by librarians was also noted by Roncevic (2013, 5) and Frederick (2015, 206).

Regarding the assignment of headings, Stachokas (2009, 208) indicated that “the not-so-hidden secret of academic libraries today is that many cataloguing departments do not catalogue electronic resources such as e-serials, databases and e-books.” On the other hand, Martin and Mundle (2010, 235) in the context of the massive incorporation of e-books and the metadata provided by the providers noted that “libraries by necessity may need to give up some control over their data as they find more efficient ways to serve their users. Tools to assist in the evaluation, clean-up, and enhancement of records in batch processes will become even more important.” We understand that this cession or loss of control extends to several dimensions, among them the metadata related to the indexing and classification of e-books. Thus, according to the data we have obtained, 35% of the librarians surveyed do not assign subject headings to e-books and 58% do not do so because they are provided by publishers/suppliers. Also, in another question, 64% expressed that publishers/suppliers should provide the subject headings as opposed to only 20% who were against. This means that the presence of headings in the bibliographic records that are so useful in the discovery phase is practically being left to the e-book providers.

Librarians indicated that they do not assign subject headings because of the instability of the collections or staff scarcity. They did assign subject headings when the e-books were of great interest, the bibliographic records provided contained no subject matter, were inadequate, more specific subject matter was desired, the e-books were part of recommended bibliographies in the subject teaching guides or were purchased individually and not part of subscribed packages. In this sense, our data are aligned with what Wu and Mitchell (2010, 169) already pointed out when they stated that librarians only catalogue e-books when they are of great value to the collection, are proprietary or the providers do not provide the bibliographic records or are of questionable quality.

Libraries have been dealing with different criteria to select providers for the subscription or purchase of e-book packages. Perhaps the most widespread are price, subject areas and access models. In EBLIDA Key Principles (2012), access, ongoing access, long-term preservation, metadata availability, price and privacy are mentioned. Vasileiou, Hartley and Rowley (2012) refer to business model, license, price, platform features and interface, thematic coverage and correspondence with reading lists. Roncevic (2013) talks about content, technical specifications, functionality and business model. Yuan et al. (2018) referring to the University of Toronto Libraries talk about a model that guides procurement decisions and practice in evidence-based approaches, user demand, print usage pattern, cost and availability. Ciptayani and Dewi (2018) present a design of a decision support system for e-book selection based on the following criteria: price, DRM, content, type of provider, business model, license, technical support, resource capacity and customer support. Finally, concerning this, we offer some words from Albanese (2007) who, when talking about a survey answered by more than five hundred libraries, most of them academic, started that the respondents revealed that price is the main concern in the purchase of e-books, followed by content.

It is noteworthy that the quality of the bibliographic records in general and of the subject headings (as a fundamental element for the location of e-books during retrieval and their subsequent use) are not essential criteria that guide the decision-making of librarians. As mentioned above, the expert group that drafted the EBLIDA Key Principles on the acquisition of and access to E-books by libraries in 2012 identified one criterion to be considered: that e-books should be viewed with metadata but without express mention of minimum quality. It has also been noted above that numerous papers have drawn attention to both the gaps and the variety of errors in the bibliographic records (Sanchez et al. 2006; Wu and Mitchell 2010, 171; Zhao and Zhao, 2010, 98; Breedt and Walter 2012, 5; Wiersma and Tovstiai 2017, 624; Mi and Pollock 2019).

We should remember here that when asked if they considered the quality of the headings and classification codes provided by the suppliers when selecting e-book packages, 54% said that they did not take this into account, compared to 35% who did. In this regard, Sanchez et al. (2006, 69) noted that “with the proliferation of e-book sources that use very basic cataloguing or none at all, we will face larger issues of how, or if, we can continue to provide consistent, quality cataloguing and authority control for these titles. If some entity does not provide cataloguing for the universe of e-books, will other methods such as basic Internet search engines be sufficient to provide access?”
As regards the use of by users it is interesting to highlight that traditionally p-books have existed because readers have them readily available on shelves and they can also search for their metadata in catalogs. Therefore, there is a double path to reach them. However, nowadays an e-book exists for a user if its metadata exist and are traceable with the discovery tools. In this sense, Yuan et al. (2018) in their study on the use of identical titles in p-books and e-books formats by users have highlighted the importance of metadata to promote the use of electronic resources and specifically noted (2018, 47) that if there were no e-book metadata within the library catalogue, e-book usage was low. Park and Tosaka (2010, 705) also wrote that the quality of metadata is essential for resource discovery and sharing. And equally, therefore, discovery tools are essential in this electronic environment.

Concerning the use of e-books by users, it is interesting to note that traditionally a p-book has existed for users because they have it available on the shelf and it also has its descriptive metadata in the catalogue. According to our data, only 48.1% of the librarians consider that the user is being offered easy and interactive discovery tools and, on the other hand, we should remember that 80% consider that the assignment of subject headings to e-books is useful for users when using the discovery tools. This in line with Ravit and Dana (2015, 809) who in a study by the Ontario Council of University Libraries’ Scholars Portal Platform found that 8.7% of users used the "subject" field to search for e-books, 11.1% used the "author" field and 19.1% used the "title." In another similar study, Ravit and Dana (2015, 810) on a set of almost fifteen thousand searches randomly selected 637 and, after their detailed analysis, found that 40% of the searches were performed on the "subject" field, 30% combining "subject or title" and 9% combining "subject or author." They wrote (2015, 812) that "analysis of our use data suggested that subject searching and topical searching by users are more important to discovery than we had originally thought." The importance that users confer to the "subject" field is, therefore, clear. Stachokas (2014, 47) highlights the subject expertise when talking about types of professional services that librarians offer in the current library (hybrid library) and will offer in the future with the electronic library. Therefore, subject headings seem to be of interest to users and are used in their searches for information, but at the same time, the process that produces subject headings is either being abandoned or is being outsourced to e-book publishers/providers.

In summary, with this massive incorporation of e-books it seems that we are heading towards the quicksand where a weak polyhedral reality with deficient or improbable metadata is installed, with a few librarians who do not have the capacity to perform an exhaustive control of the records, where e-books without quality metadata have a high probability of not being recovered and, therefore, of not being used; on the other hand, as some studies point out (Jeong 2012; Miller 2014; Perrin 2016; Fry 2018; Aharony and Barllan 2018; Bergstrom and H.glund 2018; Kristensen and Kampen 2019), there could be a possible stagnation or decline in the acquisition of p-books, even though many users seem to prefer this format.

Librarians expressed that they do not assign classification codes for the following reasons: 1) the effort does not compensate for later use; due to the fluctuation of e-book packages they do not want to do work that could be lost; 2) it is irrelevant information during the user’s search; 3) because of the high workload and that classification codes are less important in electronic management; or, 4) the search system does not offer the possibility of retrieving e-books by classification.

Finally, several authors have used the concept of disruptive technology to refer to what the emergence of e-books in libraries has meant (Dillon 2002; Lewis 2004; Lafferty and Edwards 2004; Berube 2005; Wilson 2014; Goedeken and Lawson 2015; Frederick 2016; Aggerbeck et al. 2016). This expression has also been used in relation to the emergence of open access (Lewis 2012; Moulaison and Million 2014; Tapfuma and Hoskins 2019). David W. Lewis (2004) reviewed the changes that have occurred in libraries by analyzing collections, bibliographic control and reference service based on the theories of Clayton M. Cristense (1997), who coined the term disruptive technology. The closest question we asked in our survey on this subject was whether such significant changes were taking place in relation to the technical processes performed in e-books (assignment of subject headings or classification numbers) such that we could speak of a change of paradigm or a change of model. From the answers given by the librarians, it appears that there is no agreement on this issue.

6.0 Conclusions

In the discussion section, various recapitulations have been presented when commenting on the data collected. However, we would like to emphasize here several aspects that seem significant to us. Indexing as a conceptual representation to satisfy information needs implies an indissoluble union between the processes of indexing, retrieval and use of information. In the current library ecosystem, an electronic resource exists if it is represented by metadata, so it exists through its metadata. Thus, e-books exist for library users if their descriptive metadata is searchable with discovery tools.

On the other hand, in the absence of other research that completes or deepens what has been studied here, or to contrast more intensively our data and interpretations with other studies, different significant perceptions emerge from the preceding paragraphs: 1) librarians expect e-books to ar-
rive with descriptive metadata relating to subject headings and classification codes; 2) the bibliographic records provided by publishers/providers seem to be improvable; 3) the quality of the metadata provided seems not to be considered when selecting the publishers for the purchase or subscription of e-books; 4) the discovery tools are also clearly improvable; 5) there seems to be no “frustration” or “stress” among librarians over changes in technical processes; and, 6) we do not seem to be facing a paradigm shift motivated by these issues.

With this massive incorporation of e-books, it seems that we are heading towards the quicksand where a weak polyhedral reality with deficient or improvable metadata is installed, with librarians who do not have the capacity to perform an exhaustive control of the records, where e-books without quality metadata have a high probability of not being recovered and, therefore, of not being used.

Therefore, taking as a reference some revealing data shown here and the possible connections that can be derived, we should consider whether it is appropriate for library managers to leave almost full control over bibliographic records to publishers and suppliers. We consider that we cannot renounce that e-books are always accompanied by quality metadata, because they are the key to their use by users. Therefore, as e-books are likely to increasingly displace p-books, it would be advisable that librarians grouped in lobbies, in strength and over time, that they demand from publishers and suppliers higher-quality metadata in general, and in particular, that the presence of metadata related to subject headings be an essential element in the bibliographic records and that these metadata of content representation are a reference point in the decision-making by librarians in the purchase of e-books.

Finally, we consider that this work has provided answers to various questions, but on the other hand, it also leads us to ask ourselves some questions that, perhaps, it would be interesting to address in future research; for example, to what extent do errors or gaps in the title, author or record subject headings metadata interfere with finding e-books using discovery tools and, therefore, with their end use? To what extent do searches in discovery tools with irrelevant results end up directing users to p-books available on shelves? Or how satisfied are users of my library with finding e-books after using discovery tools?

References


Lewellen, Rachel, Steven Bischof and Terry Plum. 2016. “EBL eBook Use Compared to the Use of Equivalent
Print Books and Other eResources: A University of Massachusetts Amherst – MINES for Libraries® Case Study.” Performance Measurement and Metrics 17: 150-64. doi:10.1108/PMM-04-2016-0013
Miller, Laura Newton. 2014. “Preference for Print or Electronic Book Depends on User’s Purpose for Consulting.” Evidence Based Library and Information Practice 9: 95-7. doi:10.18438/BBB891


Annex 1: Survey

Survey on the assignment of subject headings and classification to ebooks in university libraries

1. Please, mark as appropriate:

<table>
<thead>
<tr>
<th>Question</th>
<th>Yes</th>
<th>No</th>
<th>We’re on it</th>
</tr>
</thead>
<tbody>
<tr>
<td>Have you developed a procedure/policy for the management of electronic</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>documents?</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2. Mark as appropriate (1 Disagree; 5 Agree):

<table>
<thead>
<tr>
<th>Question</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>With the management that is currently being carried out with ebooks in</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>your library, ebooks are part of the library as well as printed books.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3. Does the library have a unique discovery tool for printed books, ejournals and ebooks?

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
<th>We are working on it</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4. Mark as appropriate (1 Disagree; 5 Agree):

<table>
<thead>
<tr>
<th>Question</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>The collections of ebooks are unstable and make it difficult to keep</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>the library catalogue updated.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

5. Are the publishers who supply ebooks to your library the same publishers who supply eBooks?

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
<th>Sometimes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

6. Mark as appropriate (1 Disagree; 5 Agree):

<table>
<thead>
<tr>
<th>Question</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>There is confusion among users about what ebooks are and what electronic journals are.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

7. Mark as appropriate (1 Disagree; 5 Agree):

<table>
<thead>
<tr>
<th>Question</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bearing in mind that the location of an ebook can be more difficult than a printed book because it cannot be placed on a shelf, discovery and search options should take precedence in this environment. Is the user being offered real, easy and interactive options for discovering and searching information?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
8. Has it ever been the case that a librarian has assigned subject headings or classification codes to a set of ebooks and after their incorporation into the library platform, after some time, the publisher / provider of ebooks has eliminated or modified the records in overlay processes?
[ ] Yes
[ ] No
[ ] I do not know
Observations:

9. Does your library assign subject headings to purchased or subscribed ebooks?
[ ] Yes
[ ] No.
[ ] Yes to some of them.
Please comment briefly on why or on which:

10. If the answer to QUESTION 9 has been affirmative, is the process of assigning subject headings the same as for printed books?
[ ] Yes, the same as for printed books
[ ] It is not the same.
Please comment briefly on the difference:

11. If the answer to QUESTION 9 has been NEGATIVE, check as appropriate (1 Disagree; 5 Agreed):

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>We do not assign subject headings to ebooks because they are provided by publishers/providers.</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
</tr>
</tbody>
</table>

12. Mark as appropriate in relation to metadata provided by publishers/providers (1 Disagree; 5 Agreed):

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metadata related to the subject headings of ebooks are adequate.</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
</tr>
<tr>
<td>Metadata related to subject headings are sufficient and are not worth modifying.</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
</tr>
<tr>
<td>Metadata related to the classification codes of ebooks are adequate.</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
</tr>
<tr>
<td>Metadata relating to classification codes are sufficient and are not worth modifying.</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
</tr>
</tbody>
</table>

13. Does your library assign classification codes to purchased or subscribed ebooks?
[ ] Yes
[ ] No. Please comment briefly on why
[ ] Some of them do. Please comment briefly on which
Please comment briefly on why or on which:

14. If the answer to QUESTION 13 has been affirmative, is the process of assigning classification codes the same as for printed books?
[ ] Yes, the same as for printed books
[ ] It’s not the same.
Please comment briefly on the difference:

15. If the answer to QUESTION 13 has been NEGATIVE, check as appropriate (1 Disagree; 5 Agreed):

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>We do not assign classification codes to ebooks because they are provided by publishers/providers.</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
</tr>
</tbody>
</table>
16. We want to know your opinion on the following aspects. Mark as appropriate (1 Disagree; 5 Agree):

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>The high number of ebooks purchased/subscribed in the library makes an adequate assignment of <strong>subject headings</strong> unfeasible.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The high number of ebooks purchased/subscribed in the library makes an adequate assignment of <strong>classification codes</strong> unfeasible.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ebook publishers / providers should provide the <strong>subject headings</strong>.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ebook publishers / suppliers should provide <strong>classification codes</strong>.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>In the selection of ebook packages (eBook packages) for purchase or subscription, the quality of the metadata related to the <strong>subject headings</strong> and the <strong>classification codes</strong> provided by the publishers / suppliers is taken into account.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Librarians must assign <strong>subject headings</strong> to ebooks.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Librarians must assign <strong>classification codes</strong> to ebooks.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

17. We assign more or less the same number of subject headings to ebooks as for printed books.

<table>
<thead>
<tr>
<th></th>
<th>Yes.</th>
<th>No.</th>
<th>I don’t know.</th>
</tr>
</thead>
</table>

Please comment briefly on why:

18. If there have been changes in applying the technical processes to ebooks in relation to indexing and classification, have they been introduced in the library’s procedures manual?

<table>
<thead>
<tr>
<th></th>
<th>Yes.</th>
<th>No.</th>
<th>I don’t know.</th>
</tr>
</thead>
</table>

Please comment briefly on why:

19. Do you consider that assigning subject headings to ebooks is useful for users when handling discovery tools?

<table>
<thead>
<tr>
<th></th>
<th>Yes.</th>
<th>No.</th>
<th>I don’t know.</th>
</tr>
</thead>
</table>

Please comment briefly on why:

20. Do you consider that assigning classification codes to ebooks is useful for users when handling discovery tools?

<table>
<thead>
<tr>
<th></th>
<th>Yes.</th>
<th>No.</th>
<th>I don’t know.</th>
</tr>
</thead>
</table>

Please comment briefly on why:

21. Mark as appropriate (1 Disagree; 5 Agreed)

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>I consider that there is some ‘frustration’ or ‘stress’ among the librarians responsible for executing these technical processes.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

22. If you have marked 4 or 5 in the previous question, briefly indicate why you think there is some ‘frustration’ or ‘stress’ among the librarians responsible for executing these technical processes. Please explain it briefly:
23. Mark as appropriate (1 Disagree; 5 Agreed)

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>In this new and massive digital environ-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ment, it would be appropriate to intro-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>duce the automatic assignment of sub-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ject headings.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

24. Mark as appropriate (1 Disagree; 5 Agree):

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>In this new and massive digital environ-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ment, it would be appropriate to intro-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>duce the automatic assignment of classi-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fication codes.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

25. I believe that such significant changes are taking place in relation to the processes of assigning subject headings or classification codes to ebooks that we could talk about a paradigm shift, a model change.

[ ] Yes.

[ ] No.

[ ] I do not know.

Please comment briefly on why:

---

Annex 2:
The reasons why librarians feel some “frustration” or “stress”

- The number of documents is overwhelming. In mass uploads, we even try to translate some basic or general subjects.
- This is a very time-consuming task, and there are few staff to do this.
- It is difficult to achieve the objectives and process all the e-books properly.
- The records provided by suppliers force many revisions and still break, in many cases, the consistency of the catalogue.
- It is a very difficult task to keep up to date.
- Loss of control.
- Any change in the way of operating implies a certain degree of stress until the procedures are clear. On the other hand, there is a lot of cases in the e-books (some are property, others by subscription, others large collections, others individual), and it is difficult to have a single and clear procedure, many times it is decided on the fly.
- Subjects in English, bad quality in the MARC ... instability of the collections, now they are, now they are not, especially in subscription packages.
- The quality of the original metadata of e-books is very poor, requiring a high level of reprocessing by librarians.
- Due to the more unstable character of signed digital libraries, exclusion and inclusion of titles by the provider, for example, can generate a feeling of work without completion.
- There is frustration because of the amount that is processed at once, with records provided by publishers, which necessarily interferes with the quality of records.
- Concerns exist because there are many variables to be considered in this “hybrid” phase. The normative instruments of the area do not fully cover this type of material.
- The migrations when they take place leave the catalogue inadequate for searching and the work of redoing the processes frustrate, because it seems that we are doing rework.
- Many doubts when processing, because it is a virtual exemplar.
- Because the indexing of e-book providers is very general and little specific. My impression is that this task is not performed by librarians.
- It is not always possible to reconcile the demands of analysis with the quality and depth needed to represent the contents.
- Frustration or stress comes from the fact that technology evolves, but quickly that the training of librarians and also the adjustments of software tools and hardware that librarians use are always out of step with technological evolution. The library has been the poor relative in the adequacy of both human and technological resources and then it cannot function in the timings of social evolution.
- They deal with criticism from selectors and reference when e-book vendor records are inadequate for discovery. Selectors are furious about paying for e-books with bad metadata.
- The quality of records vary from vendor to vendor, so different bulk edit operations have to be performed
based on the vendor. Some vendors only supply files of all the records rather the ones that were added/deleted since the last load. The most harrowing issue arises when vendors decide to change their unique identifiers so there’s no longer a reliable match point, e.g., ebrary and EBL merged to ProQuest e-book central.

- There is not always a valid subject heading to cover the subject matter of the material; this can lead to stress as the librarians are unsure if there are enough relevant subject terms used so that the item can be discovered.
- Because we’re importing large datasets of e-books, it then becomes difficult to check the quality of the bibliographic data.
- Often requires complex technical workflows to ensure quality of records and correct coverage of subscriptions in supplied record sets.
- There are not enough staff to do this properly, and yet it does need to be done properly, putting a lot of pressure on existing staff. Non-cataloguers do not seem to understand this and are unhelpful and yet are the first to complain if an e-book is not properly catalogued.
- The technical processes are complex
- Knowing we should, but do not have the time, would create stress. We do not provide the right metadata to make our e-books found by the users. This is especially the case as we do not use English as our primary language, and subject headings/key words assigned to our books will be in our maternal language for p-books.
- The county keeps on changing the system that delivers the e-books. New systems all the time
- We do not have resources to catalogue all e-books in our catalogue, but we activate them in our link resolver. The metadata in the link resolver is very limited, and unfortunately it cannot be enriched even manually. Not being able to let the users find books by subject headings given by the library can cause stress for the librarians wanting the purchased resources to be found and used by our patrons.
- Some vendors do not employ qualified cataloguers to create the data, so it is often lacking, substandard or inadequate.
- I believe the subject heading and classification codes might be helpful, but we do not consider these as major factors in e-book discovery. We rely on the deep metadata behind the discovery layer and have seen the increase of usage in e-books without local work to update title records
- Staff time and knowledge in subject heading and classification codes are limited.
- The quality of metadata from each supplier varies from average to lamentable. Each supplier’s records needs a different process to massage them into a form acceptable to us.
- Frustration is caused by poor metadata in catalogue records provided by providers. Unless we can obtain records from another source (e.g., OCLC’s WorldShare) there is little we can do to rectify the problem, apart from try to work with the provider for them to improve the records.
Depicting Historical Persons and Identities: A Faceted Approach†

Brian Dobreski*, Jian Qin**, and Melissa Resnick***

*University of Tennessee, Knoxville, School of Information Sciences, 449 Communications Bldg, 1345 Circle Park Dr., Knoxville, TN 37996, <bdobreski@utk.edu>

**Syracuse University, School of Information Studies, 311 Hinds Hall, Syracuse, NY 13244, <jqin@syr.edu>

***University at Buffalo, Jacobs School of Medicine and Biomedical Sciences, 955 Main St., Buffalo, NY 14203, <melissa_resnick@hotmail.com>

Brian Dobreski is Assistant Professor in the School of Information Sciences at University of Tennessee-Knoxville. His research focuses on the social implications of metadata, resource description, and other knowledge organization practices, as well as the concepts of personhood and personal identity. Brian received his PhD in information science from Syracuse University. His dissertation, ‘Values in Knowledge Organization Standards: A Value Analysis of RDA,’ received an honorable mention for the ProQuest Best Dissertation Award, as well as runner up for the iSchools 2020 Doctoral Dissertation Award.

Jian Qin is Professor at the School of Information Studies, Syracuse University. Her research areas include metadata and knowledge modeling, knowledge networks in biomedical data, research data management, and scholarly communication. She has published widely in library and information science journals and given presentations at numerous national and international conferences and workshops. Her research has been funded by National Science Foundation and Institute for Museum and Library Services among others. She is the recipient of the 2020 Frederick G. Kilgour Award for Research in Library and Information Technology. Qin holds a PhD degree from University of Illinois at Urbana-Champaign.

Melissa Resnick is currently a post-doctoral fellow at the University at Buffalo, Jacobs School of Medicine, Department of Biomedical Informatics. She holds a PhD from the University of Texas Health Science Center at Houston, a MLS from the City University of New York's Queens College, a MS from Rensselaer Polytechnic Institute, and a BA in biology from the University at Albany. Her doctoral work was in classifying healthcare providers' information needs using Medical Subject Headings (MeSH). Her research interests include terminologies and their use in knowledge organization systems.


Abstract: Archives are responsible for presenting historical materials to users while also placing them in context. Historical individuals and their identities pose specific challenges for the archive, including how to negotiate changing cultural perspectives on identity and how to convey and explain contexts to modern audiences. Contemporary subject representation practices in cultural heritage tend to offer a reductive view of personal identity. Using a collection of visual images of nineteenth and twentieth century sideshow performers, the present work explored a faceted approach to representing personal identity in order to: 1) offer a more holistic view of historical individuals, their identities, and relationships; and 2) help place historical terms and perspectives into context in order to better support user understanding. Utilizing an inductive approach, the researchers developed a framework of seven facets of personal identity. A textual keyword analysis of existing title metadata for images in the collection was implemented in order to identify distinct individuals depicted and any facets associated with them. Finally, collection-based language and closest matches from several controlled vocabularies were used to assign values to the facets for all individuals. The resulting metadata offers a flexible, richer, more holistic analysis of personal identity for these materials, while also exposing some deeper issues concerning identity, performance, and visual imagery. Opportunities exist for further automation and application of this faceted system to other historical collections.
1.0 Introduction

Advances in collection management, preservation, digitization, metadata, and data publishing have improved access to historical materials and brought them to new and expanding audiences. As stewards of these materials, libraries, archives, museums, and other cultural heritage institutions are responsible not only for presenting them to users but also for helping the users understand and make sense of them as well (Society of American Archivists 2013). Conveying and explaining the context around historical materials, including the persons, places, groups, and terminologies associated with them, is thus a critical task within the cultural heritage domain. This is also a difficult task, however, as changes in conceptualization, language, and perspective over time can leave contemporary audiences either unable to discover what they need or confused or offended by what they may find. This problem is particularly apparent in the subject metadata associated with historical materials. How can subject representation address historical materials that have been seen through changing contexts and perspectives over time? How can subject languages describe resources from a variety of times and places while accommodating modern vocabularies? And, by labelling older materials with contemporary terminologies and perspectives, do cultural heritage institutions actually run the risk of obscuring context and misrepresenting these resources and their original cultural settings?

These problems become particularly challenging when dealing with resources that depict historical persons. Recording and using metadata about persons and their identities is a complex but significant area of work in cultural heritage (Dobreski and Kwaśnik 2018), one that becomes even more complicated when representing persons who have been viewed from sharply changing perspectives over time. One group of persons that exemplifies such challenges is American sideshow performers. Sideshows or freak shows were widely accepted forms of entertainment in the United States from mid-nineteenth to mid-twentieth centuries, showcasing individuals with physical abnormalities and promoted through sensational imagery and language. In the twentieth century, public opinion shifted from spectacle and curiosity to embarrassment and sympathy as medical science was able to explain performer conditions in scientific terms (Bogdan 1988). Over time, popular and controllable vocabularies for the conditions these performers embody have grown more reflective of the medical and scientific community’s perspectives (Adler et al. 2017). In describing materials that depict these persons, modern subject representation thus runs the risk of reducing historical performers to their clinical conditions and obscuring other important context concerning their identities and lives. In a review of subject analytic approaches to such materials, Rinn (2018) argued that a range of diverse terminology should be used to capture the whole of a person’s identity and experience but found that accomplishing this through traditional subject representation strategies was difficult. Indeed, work has shown that traditional cultural heritage metadata generally lacks the capability to fully capture contemporarily relevant aspects of identity, such as race, ethnicity, and culture (Clarke and Schoonmaker 2019). Facet analysis, a longstanding approach to providing a flexible, multi-perspective representation of subject content (La Barre 2010), has been underutilized in this area but may offer one means of addressing these issues.

In this study, an exploration of faceted subject representation for historical depictions of persons is presented through a case study of the Ronald G. Becker Collection of Charles Eisenmann Photographs. Housed in the special collections of Syracuse University, the Becker-Eisenmann materials constitute an archival collection containing over 1,400 images of sideshow performers from the nineteenth and twentieth centuries. Originally intended for promotion and entertainment, these images now serve as an intriguing but complex set of primary sources for users interested in the areas of history, disability, culture, and performance. Previous work with the Becker-Eisenmann and similar collections has shown current subject analysis to be reductive while also struggling to navigate depictions of disability and outdated or offensive language (Rinn 2018; Dobreski et al. 2019). The present work proposes a faceted approach in order to: 1) offer a more holistic view of these performers, their identities, and relationships; and 2) help put different perspectives and terms into context to help users better understand these materials. This is realized through a multipart analysis designed to determine a framework of useful and easily available facets related to the depicted persons, followed by a demonstration of how a mixture of vocabularies may be used within this framework to accomplish a more complete subject representation. This approach holds
potential for the Becker-Eisenmann and similar collections, and stands to help their users through offering a more comprehensive treatment of the individuals and complex identities depicted by historical images.

2.0 Background and literature

Though cultural heritage institutions are responsible for creating and maintaining metadata about information resources such as documents, works of art, and artifacts, many are responsible for recording information about persons as well. Representing persons including authors, interviewees, or biography subjects is a significant area of work in cultural heritage knowledge organization and requires the collection and presentation of metadata about persons and their identities (Dobreski and Kwaśniki 2018). Representing personal identities bears distinct challenges concerning terminology and labelling practices though, and the cultural heritage domain has been criticized for failing to capture contemporarily relevant aspects of identity, such as race, ethnicity, and culture (Clarke and Schoonmaker 2019; Wright 2019).

These challenges and failures are particularly apparent in relation to collections of materials associated with American sideshows. Sideshows, or freakshows, were a culturally accepted entertainment form in the nineteenth and twentieth century in which persons with unusual characteristics performed for or were displayed to paying spectators. Though some persons were unusual due to their abilities (e.g., fire eating), more were found unusual due to mental or physical differences that would be considered disabilities in modern culture (Kochanek 1997). Many of these individuals were taken from their families under false pretenses with little information remaining about their real identities and origins (Mateen and Boes 2010). In marketing and promoting these performers, sideshow managers often developed stage personas for them involving elaborate, exotic, and often fraudulent backstories that further obscured their identities (Bogdan 2012). Twentieth century shifts in public sentiment and acceptance of sideshows may be attributed to the growth of modern medical science which changed the conception of sideshow performers from unique spectacles to persons afflicted with documented medical conditions (Kochanek 1997). Sideshows waned in popularity as the American public came to view them as unsympathetic and exploitative. In organizing historical materials associated with these sideshows, contemporary cultural heritage institutions face the significant challenge of accurately and respectfully representing sideshow performers and their identities while allowing users to understand the changing cultural milieus in which they have been situated.

One of the more common types of cultural heritage collections associated with sideshows are photographic collections of performers. These photographs were originally sold to generate revenue and advertise for sideshow performances and were typically composed in such a way as to aggrandize or exoticize the depicted performer through spectacular costumes, posing, and evocative titles (Bogdan 2012). Such depictions were inherently exclusionary and marginalizing and pose challenges to the representation, organization, and discovery of these materials (Altermark and Edenborg 2018). In reviewing descriptions from several collections of sideshow performer photographs, Rinn (2018) found that contemporary subject representation of these persons tended to reduce them to medical diagnoses while neglecting other aspects of their identities. Indeed, Adler et al. (2017) found that common cultural heritage subject languages such as LCSH have developed in such a way as to reflect the medical community’s perspectives while ignoring other perspectives and terminologies associated with disabilities and differences. Representing sideshow performers in the cultural heritage domain is thus often reduced to a task of retrospective diagnosis, labelling historical persons with contemporary medical conditions in spite of incomplete diagnostic information (Mateen and Boes 2010). In studying the research activities of disabilities scholars, Koford (2014) found that many of them did not utilize subject headings during their search process, citing differences in terminological usage and the presence of offensive or confusing labels. Accordingly, Koford called for knowledge organizers to pursue new tactics for subject representation for these materials. Other researchers have similarly called attention to the ways in which current subject representation in cultural heritage settings has continued to reduce and marginalize people with disabilities and physical and mental differences, suggesting alternative, more inclusive strategies be pursued (Beattie 1997; Adler et al. 2017). While making a similar call, Wright (2019) at the same time cautioned that cultural heritage institutions should continue to acknowledge historical language in a manner that explains and provides context without reproducing historical, oppressive power structures. Cultural heritage institutions are thus tasked with a number of challenges in representing, explaining, and providing access to materials that depict sideshow performers, challenges that may require strategies beyond current subject representation strategies.

One alternative approach to subject representation and access that holds promise in this area is faceted classification, a long-standing means of providing multi-perspective representations of content (La Barre 2010). In a faceted representation system, subject content is divided amongst a specific set of conceptual categories and applied across the resources being organized. Facets can thus be thought of as recurring characteristics that are meaningful to some set of users, and different sets of materials may be analyzed along different sets of facets (Hudson 2019). As an organizational structure, facets can be traced back to Aristotle’s categories,
and more recently, the works of Ranganathan and Bliss (Mills 2004). Though they have traditionally been considered a rationalist approach to organizing, based on the a priori knowledge of the organizer (Hjørland 2013), facets are not entirely without empirical basis. During the process of facet analysis, terms in a given corpus may be examined in order to determine what general concepts they represent; from this, general classes of concepts and relationships emerge and may be used as the basis for facets (La Barre 2010). Facets have become an increasingly common classificatory feature in many information retrieval tools in web environments, with users growing more accustomed to them, particularly in commercial settings (Broughton 2006). In these interfaces, facets often manifest in the form of search filters, designed to help users narrow down large result sets. Facets are a common feature in cultural heritage information retrieval tools as well, and though many such systems utilize facets to represent more descriptive features of materials, they can be used to offer robust subject representation as well. For example, Broughton (2002) explored the use of faceted representation for a collection of digital objects whose current DDC and LCSH subject representation was not satisfactory for conveying the collection’s complex subject matter. Utilizing vocabulary analysis, the study resulted in the construction and application of subject facets capable of providing more meaningful access for digital humanities researchers. Facets can thus offer an alternative and less reductive means of conveying subject matter and terminology for complex historical materials.

### 3.0 Methodology

Representing the content of photographs or images may be approached in two ways: text-based and image-based. The text-based approach uses keywords to understand and describe aboutness and of-ness, and most work in this approach is performed by human catalogers or taggers. As such, the text-based approach may bear biases due to differences in interpretation and use of terms. The image-based approach is also called query-by-image content (QBIC) that includes three types: spatial-index-based method, feature-based indexing methods, and object- and pictorial-based methods (Ewald 2009). The QBIC methods employ automatic algorithms to analyze features such as the location of an object or scene, color, texture, or statistic of a scene in the image, or symbolic image of an array representing the spatial relations among objects. Although the QBIC approach has its advantages, it also has its limitations, e.g., inability to detect and understand nuanced historical and social aspects of image content. Given the pros and cons of both image indexing approaches and the existing textual data in the source collection, we decided to adopt the text-based approach; we acknowledge, however, that future QBIC-based approaches may hold potentials for these materials as well.

The data used for this project included the Becker-Eisenman Collection of digitized photographs and the associated item-level metadata. The data set contained 1,415 records representing unique images held in the collection. These metadata records utilize a modified version of Dublin Core. Current subject representation for these materials consists of a single LCSH heading per record. Despite the presence of subject metadata, as well as brief descriptions for each image, we chose to limit our analysis solely to the title element. Title metadata for the images reflects the monikers printed on them at the time of their publication; in the absence of a printed title, archivists had previously devised and assigned a brief title using language from similar objects in the collection (e.g., “Unidentified Bearded Lady”). Data in the title element was thus taken to be the most “primary” and contemporaneous to these materials while being least interpreted or translated by subsequent audiences as subject and description elements may be.

Within the collection, multiple images sometimes shared the same title (e.g., four different photographs with the title “Nora Hildebrandt, Tattooed Lady”). As a result, for the 1,415 records, 957 distinct titles were determined and collected. Using this title data, we devised a multi-part methodology designed to detect facet patterns relating to depicted persons and use these facets to enrich the representation of historical and social aspects of their identities. Analysis proceeded in three distinct phases: facet determination, facet analysis, and facet assignment.

#### 3.1 Facet determination

Prior to full analysis of the collection metadata, the researchers worked to determine the set of available facets relevant to the materials as well as to the goals of the study. During this phase, we reviewed sample materials and their title metadata in order to identify recurring aspects of personal identity. During this process, additional sources of reference for describing persons and historical background about persons were also consulted, including the Library of Congress Subject Headings (LCSH), which contains many headings relating to groups of persons, Resource Description & Access (RDA), a descriptive standard that includes instructions on creating personal authority records, and Wikipedia from which we found articles on the lives of some of the sideshow performers. Within these standards and reference sources, certain aspects of identity were common, including gender, ethnicity, age group, language, occupation, and relationship to other persons and groups. The researchers used these standards and sources as a point of reference in their determination of facets. The overall process was inductive and iterative, featuring a recurring process of group
3.2 Facet analysis

With the facets of interest determined, we performed descriptive analysis on the title metadata for the 957 unique image titles in the collection. Manual keyword analysis and text extraction was performed on titles to isolate descriptive terms and phrases associated with persons in each of the images (e.g., “Chinese midget,” “Bearded lady”). During keyword analysis, researchers focused on two tasks: 1) determining how many distinct persons were present; and 2) determining how many facets could be identified for each of those persons.

To determine the number of persons present in each image, titles were reviewed for keywords and grammatical expressions signaling the presence of distinct persons. In addition to personal names, linguistic features in title phrases provided hints for distinguishing between individual persons. In the original titles in Table 1, for example, punctuation (commas) and symbols (dashes) served as cues for differentiating one person from another. Conjunctions and prepositions were also used as cues to distinguish individual persons. For example, for the title, “Major Little Finger and Wife, midgets,” two persons were noted as being present. Wording in some titles made the determination of number of persons difficult. For the title, “Italian midgets and family,” no distinct persons could be determined through key-

To determine the number of facets present in each image, titles were reviewed for keywords and phrases associated with each of these persons, including age, gender, race, family relationship, and other facets established during the first phase of work; keywords and phrases were grouped into one or more of these facets where possible. Family relationships were indicated by keywords such as “wife,” “sister/sisters,” and “mother.” Keywords indicating the gender facet included “boy,” “girl,” “woman/lady,” and “man/men.” Forenames and titles were also seen to indicate gender for some persons. Since exact ages were not given, keywords such as “woman/lady,” “man/men,” and “mother” were taken to indicate an adult age group, while those such as “child,” “boy,” and “girl” were indicative of a child. Keywords representing nationality and race encompassed such terms as “Borneo,” “Chinese,” “German,” and “Black.” Keywords associated with the role facet described a type of performance the person was engaged in, such as “snake charmer” or “dancer.”

Single keywords in the titles were often enough to identify the age, gender, race, and role facets for individual persons shown in the photos, but in some instances, single words were too ambiguous to clearly signal the presence of one of the seven facets (e.g., “broken,” “sacred,” “frizzy”). For single keywords where no facets were clearly indicated, proceeding or succeeding keywords were appended to form a phrase. Marks of punctuation and syntactical clues were used in order to determine whether to check the proceeding or succeeding word. The resulting phrases were then reviewed to see if a facet could be determined. In some cases, even phrases were unable to indicate a facet (e.g., “Unidentified person”). Table 1 provides an example of the in-depth, keyword and phrase analysis of several image titles.

Throughout this process, a dictionary of all distinct keywords and phrases, along with the facets they were associated with, was developed. This would be utilized in the subsequent methodological steps.

3.3 Facet assignment

With facets determined for each person indicated by each title in the collection, the next phase involved assigning keywords into each of these facets. The first task here involved the assignment of collection-based, natural language keywords into facets for all human entities depicted by the items. In many cases, this facet assignment was completed during the facet analysis process described above. In such cases a relevant keyword or phrase from the title that indicated the presence of a facet could be assigned directly into that facet, for example, “giant” into the condition facet, or “black” into the race facet. For other titles, however, a keyword or phrase only implied the value for a facet, for example “Lord” implying a “male” value for the gender facet. In some instances, a keyword had both direct and implied facet values, for example “wife” being directly usable in the relation facet while also implying “female” for the gender facet.

To assign the implied facet values for all human entities represented within the collection, researchers began with a manual review process. Keywords such as “Lord” were reviewed and mapped to value “male.” In determining appropriate values, we attempted to use keyword forms already present in titles in the collection where possible. As this mapping process continued, the keyword dictionary was updated to include these mappings, allowing for semi-auto-
mated mapping of previously encountered terms (e.g., all instances of “Lord” were assigned a value of “male” for the gender facet). At the conclusion of this process, all entities had collection-based, natural language keywords assigned to all of their facets.

In the second task of the facet assignment phase, researchers turned to additional terminological sources to assign alternative values to each facet, thereby depicting multiple perspectives. Library of Congress Subject Headings (LCSH), SNOMED Clinical Terms (SNOMED CT), and Wikipedia were chosen due to their availability, applicability to collection materials, and ability to represent a range of perspectives. Commonly applied to cultural heritage materials, LCSH offers a general controlled vocabulary accommodating a wide range of concepts. SNOMED CT offers a distinct controlled vocabulary reflective of contemporary medical perspectives. While not a formal controlled vocabulary, Wikipedia offers a modern, popular perspective on terminologies; researchers relied on the article titles as representations of the “preferred” terms within this source.

For each facet value, researchers manually mapped the collection-based, natural language keyword value to its closest match within each of the three sources. In instances where a concept had no associated term or close match in a particular source, no facet value as assigned from that source. As in the first task, as manual mapping proceeded the keyword dictionary was updated, recording the chosen mappings for concepts encountered.

4.0 Results

4.1 Facet determination

The facet determination process resulted in seven recurring facets that researchers found to be useful in conveying relevant aspects of personal identity within the Becker-Eisenmann Collection. Table 2 lists these seven facets, along with example textual indicators from title metadata and sample facet values. These specific facets were used as the framework for facet analysis and assignment in subsequent phases of research. Though additional facets such as time period may be derived from further knowledge about the collection, the current study remains limited to only those facets that could be determined from title metadata.

<table>
<thead>
<tr>
<th>Facet</th>
<th>Example Textual Indicator</th>
<th>Example Facet Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Siamese twin babies</td>
<td>babies</td>
</tr>
<tr>
<td>Condition</td>
<td>Che-Mah, Chinese Dwarf</td>
<td>dwarf</td>
</tr>
<tr>
<td>Gender</td>
<td>Rose Leslie, Fat Lady</td>
<td>lady</td>
</tr>
<tr>
<td>Nationality</td>
<td>A German Woman</td>
<td>German</td>
</tr>
<tr>
<td>Race</td>
<td>Thomas Brockman, black giant</td>
<td>black</td>
</tr>
<tr>
<td>Relation</td>
<td>Pat O’Brien and Wife, Giants</td>
<td>wife</td>
</tr>
<tr>
<td>Role</td>
<td>Zoe Zobedia, Snake Charmer</td>
<td>snake charmer</td>
</tr>
</tbody>
</table>

Table 2. Facets of personal identity.
4.2 Facet analysis

Out of the 957 unique titles in the collection, 800 of these were found to contain one or more distinct persons. These 800 titles represent separate 1,219 images in the Becker-Eisenmann Collection. Within these 800 titles, 1,004 persons were identified. No attempts were made to deduplicate recurring persons across titles, meaning that some individuals are represented multiple times due to their association with multiple, different titles.

For these 1,004 persons, a total of 2,245 facet values were determined. The number of facets per person ranged from zero to four, with an average of 2.23 facets per person identified. Gender, condition, and age were the most frequently identified facets. Table 3 shows the frequency of facet occurrence among the 1,004 depicted persons.

As also shown in Table 3, the facets varied in terms of how many distinct keywords and phrases they were associated with. While the age facet was identified 460 times, only forty different keywords and phrases made up these 460 occurrences. In contrast, the gender facet was made up of 381 distinct keywords and phrases; much of this may be attributed to the large number of personal forenames that could be reliably associated with a gender (e.g., “Mary,” “John”).

It should be noted that some distinct keywords and phrases appeared in multiple facets. For example, in the context of this study, “wife” is a combination of both gender and relation, and thus is associated with values for both of these facets. In fact, of the 630 total distinct keywords and phrases included in the dictionary at the conclusion of this phase, 206 were associated with more than one facet. The presence of such compound concepts is reflective of the nature of the analytical depth of the framework employed here; a larger number of facets would be expected to reveal additional compounds.

4.3 Facet assignment

The facet assignment phase of this project saw researchers using different terminological sources to assign facet values for the persons represented within the Becker-Eisenmann Collection. The first step of this phase involved the assignment of collection-based, natural language keywords into facets for all human entities depicted by the items. To accomplish this, we utilized the dictionary of all extracted keywords and the facets to which they corresponded. As such, much of the data needed for this step was already available at the conclusion of the facet analysis process described above.

Keywords in the dictionary corresponded to values in one of two ways. The first was where keywords were directly usable as facet values, i.e., overt facets. For example, “hairy” was a valid value for the condition facet, while “Sweden” was a valid value for the nationality facet. The second was where keywords were not directly usable as a facet value, but implied one, hence a “covert facet.” For example, “enchantress” implied a value of “female” for the gender facet. Some keywords fell under both scenarios. While the term “father” signified a valid value for the relation facet, it also implied an “adult” value for age, and a “male” value for gender. All keywords were manually reviewed again in this phase for both overt and covert facet values.

Example terms from the keyword/facet dictionary are depicted in Table 4, with covert facet values denoted by the use of square brackets. “Girl” and “Princess Pontus” offer examples of phrases associated with multiple facets.

At the conclusion of this process, the dictionary of 630 keywords and phrases contained 445 directly usable facet values and 422 covert, implied facet values.

With the dictionary of collection-based language established, it was then used to assign facet values to each title, for each person depicted by the title. Title keywords were matched with keywords in the dictionary for each identified person, with the corresponding facet values then assigned. Table 5 shows the results of this process for three example titles, depicting a total of six persons.

In the second part of this phase, we turned to additional terminological sources to demonstrate how other perspectives, beyond those represented by collection-based language, could be added into our facet system. For each keyword in the dictionary, researchers reviewed LCSH, SNOMED CT, and

<table>
<thead>
<tr>
<th>Facet</th>
<th>Total Occurrences</th>
<th>Distinct Keywords/Phrases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>460</td>
<td>40</td>
</tr>
<tr>
<td>Condition</td>
<td>629</td>
<td>149</td>
</tr>
<tr>
<td>Gender</td>
<td>852</td>
<td>381</td>
</tr>
<tr>
<td>Nationality</td>
<td>45</td>
<td>24</td>
</tr>
<tr>
<td>Race</td>
<td>13</td>
<td>5</td>
</tr>
<tr>
<td>Relation</td>
<td>139</td>
<td>17</td>
</tr>
<tr>
<td>Role</td>
<td>105</td>
<td>54</td>
</tr>
</tbody>
</table>

Table 3. Facet occurrence and related, distinct keywords and phrases.
Wikipedia for closest matches for all corresponding facet values, both overt and covert. For example, for the keyword “father,” matches were determined for the values “father,” “adult,” and “male.” Only authorized terms from LCSH and SNOMED CT were considered, while article titles in Wikipedia were taken to represent the authorized term forms in this source. In cases where multiple matches existed in a single source, multiple terms were recorded. If no suitable matches were present, no terms were recorded. At the conclusion of this process, the keyword/facet dictionary was enhanced with a total of 690 terms: 249 from LCSH, 188 from SNOMED CT, and 253 from Wikipedia. The enhanced dictionary thus took the form of a crosswalk, with values from each of the three controlled vocabularies matched to the value from the original item.

With terminology from multiple sources correlated to the keywords and their facet values, researchers were able to enhance the facet metadata to include terms beyond the collection-based language depicted in Table 5. Table 6 shows one of the example titles, depicting two persons. For each person, four facets were assigned (condition, gender, relation, nationality), and values for these facets were supplied from collection-based language, LCSH, SNOMED CT, and Wikipedia.

This facet assignment test followed a workflow in which a number of steps were taken to extract facet-bearing keywords from collection-based terminology, assign keywords to facets, and normalize keywords under each facet based on multiple controlled vocabularies. The fact that multiple vocabularies were used in facet assignment creates a novel ap-

<table>
<thead>
<tr>
<th>Keyword</th>
<th>Facets</th>
<th>Age</th>
<th>Condition</th>
<th>Gender</th>
<th>Nationality</th>
<th>Race</th>
<th>Relation</th>
<th>Role</th>
</tr>
</thead>
<tbody>
<tr>
<td>Major</td>
<td></td>
<td></td>
<td>[male]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Midgets</td>
<td>midget</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chinese</td>
<td></td>
<td></td>
<td></td>
<td>Chinese</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Couple</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>couple</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Girl</td>
<td>[child]</td>
<td></td>
<td></td>
<td>girl</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Princess Pontus</td>
<td>[female]</td>
<td></td>
<td></td>
<td></td>
<td>Princess</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4. Examples from dictionary of keywords and corresponding facet values.

<table>
<thead>
<tr>
<th>Human Entity</th>
<th>Cumulative Facets</th>
<th>Facet Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Admiral Dot ... Midgets</td>
<td>condition, gender</td>
<td>condition: midget gender: male</td>
</tr>
<tr>
<td>Major Atom ... Midgets</td>
<td>condition, gender</td>
<td>condition: midget gender: male</td>
</tr>
<tr>
<td>Unidentified Normal Sized Man</td>
<td>age, gender</td>
<td>age: adult gender: male</td>
</tr>
<tr>
<td>Moung Phoset ... Hairy Family ... Burma</td>
<td>condition, gender, relation, nationality</td>
<td>condition: hairy gender: female relation: family nationality: Burma</td>
</tr>
<tr>
<td>Mah Phoon ... Hairy Family ... Burma</td>
<td>condition, gender, relation, nationality</td>
<td>condition: hairy gender: female relation: family nationality: Burma</td>
</tr>
<tr>
<td>Original title: Thomas Brockman, black giant</td>
<td>race, condition, gender</td>
<td>race: black condition: giant gender: male</td>
</tr>
<tr>
<td>Thomas Brockman, black giant</td>
<td>race, condition, gender</td>
<td>race: black condition: giant gender: male</td>
</tr>
</tbody>
</table>

Table 5. Examples of facet values assigned with collection-based terminology.
proach to deep-subject representation. Both overt and covert facet assignment enrich the semantic breadth and depth in representing historical materials which can enable sophisticated, facet-based retrieval and increase the discoverability for these materials.

5.0 Discussion

As described above, a faceted approach to subject representation in the Becker-Eisenmann Collection allowed for a more holistic view of the persons depicted, their identities, and their relationships. By drawing out data such as gender, nationality, and family relationships, we were able to provide richer, clearer insight into aspects of personal lives and identities that are often obscured or overlooked in more traditional approaches to cultural heritage metadata. Beyond representing these persons more fully, the availability of facets could provide new search and retrieval options for users in online interfaces, many of whom are already familiar with facets and filters in contemporary web design (Broughton 2006). These facets, in the context of an information retrieval system, would allow users to ask more complex questions about these materials and the performers associated with them.

The facet determination and analysis employed in this study leveraged existing metadata to add further context around these historical materials. While the descriptive, evocative title metadata for these materials provided useful insight, use of title metadata alone is somewhat limiting and does not represent everything the archive knows about these materials. Metadata from the description field could prove particularly useful, though its analysis would require new considerations and analytic rules. Other facets may be drawn out from knowledge of the collection as a whole and its provenance. For example, time period and occupation represent additional facets that, while not explicit in the metadata, could be inferred from the materials. Further visual inspection of the images, though labor-intensive and infeasible for many large collections, could provide further insight as well. This could be particularly helpful in clarifying family relations. Regardless, the procedures of this study show the potential of enriching subject metadata through the use of pre-existing metadata even without re-analysis of each item.

It must be stated that the set of facets derived in this study unavoidably reflect the modern, academic perspectives of the researchers. Additional work should be undertaken to understand archive users’ notions of personal identities and which facets of identity are meaningful to their research and use of historical materials. The determination of values for the facets used in this study was also done from a modern perspective on the past, leading to a certain level of interpretation on the part of the researchers. For instance, terms like “Major” were taken to indicate male values for the gender facet, an assumption that may hold true for late nineteenth century materials but not other contexts. Additionally, some terms must be taken as historical euphemisms rather than at face value; for example, “baby” was a term used to imply obesity. Ultimately, whatever framework of facets is developed and implemented today may not match with future assumptions and understandings, rendering these subject representations less meaningful to future audiences. It may be worth further consideration of how to anticipate the need for materials to be reclassified by future historians who must always endeavor to interpret the past from their present milieu (Jenkins 1991).

In addition to the use of facets, this study also employed the application of multiple terminological sources in order to show variations on perspective and labelling in regards to these materials. This strategy shows potential for providing richer insight into historical identities and helping modern audiences place individuals in cultural context. It also helps mitigate some of the reductivist subject representations that have been noted in analyses of sideshow related collections.
(Rinn 2018). In the present study, we chose LCSH, SNOMED CT, and Wikipedia as additional terminological sources, though any number of controlled or uncontrolled vocabularies could be utilized. As can be seen above, the three sources we chose showed little variation overall in the terminology applied to gender and relation facets, though further variation was notable in facets such as condition and nationality. While sources such as SNOMED CT offer a deep, fine-grained approach to labelling medical conditions, this can be difficult to leverage in practice as conditions and their underlying pathophysiology cannot be fully assessed from the images and metadata. This can paradoxically result in closest matches to more general terms in SNOMED CT than in LCSH (e.g., “short stature disorder” vs. “dwarfism”). Variations in nationality terminology often reflected historical and ongoing geopolitical conflicts (e.g., “Burma” vs. “Myanmar”).

In both of these cases, the individuals depicted in this collection may be labelled with non-contemporaneous terms for their identities. The use of the original, collection-based language should thus be maintained as a means of helping users understand cultural and historical differences between contemporary and past perspectives. While Table 5 gives an example of what this might look like, care must be given to the presentation of this metadata in order to support user understanding. The original, collection-based language employs obscure historical euphemisms as well as terms found offensive by contemporary audiences, and further thought must be given toward how to utilize this language while acknowledging the marginalizing, exclusionary milieu under which it was employed. Differences among the multiple vocabularies may also result in user confusion, though providing background information on the terminological sources, their perspectives, and the reasons behind employing them may help in this regard. Any terminological source harbors its own potential biases (Knowlton 2005); transparency may be the best means of helping users navigate these. Overall, however, we believe the use of contemporaneous and multiple controlled vocabularies will be beneficial to the user. As Svenonius (2000) observes, much knowledge organization work is intended to bridge the language of resources with the language of users. In this study, we built multiple bridges from multiple facets of identity to help modern users connect with historical materials.

While a language-based approach to analyzing and representing the subjects of this collection affords a rich, flexible means of depicting complex identities, it also raises further issues around language, performance, and reality. With some exception regarding stage names, the current study takes title language literally and truthfully, though given the inherently performative nature of these materials, this is not always the case. Many individuals performing in sideshows were presented and marketed with fabricated origin stories, obscuring their original identities and the sometimes fraudulent means by which they were separated from their earlier lives (Mateen and Boes 2010; Bogdan 2012). Within the Becker-Eisenmann Collection are instances of unrelated persons depicted as siblings or spouses due to their similar conditions or appearance. Similarly, some conditions were completely “performed” by the persons being depicted, with the performers not, in fact, being conjoined twins or intersex persons. The literal approach taken in this study could undoubtedly perpetuate such misrepresentations, many of which are marginalizing and harmful in nature. Given the information we have, however, how are we to determine what was performance and what was real? For all of the work undertaken with the Becker-Eisenmann Collection, many individuals’ real identities, as opposed to their performed identities, remain obscured; our hope is that increased subject access to these materials can prompt renewed interest in this area and facilitate work being done by historians and researchers that may shed further light on the matter.

This dilemma also points to a deeper issue related to the analysis of persons and images. For instance, if a photograph is taken of a person at a costume party during which they are dressed as a doctor, the image does indeed depict a doctor, but should that person be labelled as a doctor when they are simply “performing” one? In some instances, pictures are themselves performances, not accurate depictions of individuals, and in these cases, some subject terms are perhaps more associated with the image/performance itself rather than the person. Assuming we can pick apart which images are intended to be truthful depictions of reality and which are performances, how would we convey this to users? The use of genre and form headings to denote performance or fiction may be worth further consideration in this regard. Just as films are labelled with fictive genre headings, the images could be labelled with similar terminology in order to help users more clearly understand the materials and prevent the ontological collapse of the person and the performance.

Though the overall approach and procedures described here are most immediately applicable to other collections of sideshow performer images, we believe there is potential for generalizability to other types of collections as well. Given the language-based nature of the present analysis, collections of textual materials concerning sideshow performers could benefit quite directly from similar procedures. Collections with more textual data would likely yield greater context about these performers and could further enhance our framework as well. Other collections of visual images depicting persons (e.g., postcards, portraits, carte de visite, family albums) could also be enhanced through an application of faceted subject representation of personal identity. It should be noted, however, that some level of interpretation...
was required of the researchers during the present keyword analysis, and the decisions made here may hinder generalizability to less closely related collections of materials. Also, images in the Becker-Eisenmann Collection carried descriptive, evocative titles that were useful in a large-scale reanalysis, though many other collections of visual images may not. Using textual data rather than visual inspection in our analysis allowed us to leverage existing metadata and proceed more efficiently than revisiting over 1,400 images individually. While metadata beyond the title field could be used in the analysis of other collections, there may be some image collections where further subject analysis is only possible through visual inspection. In such cases, query-by-image content (QBIC) (Ewald 2009) may prove useful, though further development of such methods would likely be necessary in order to capture the social and cultural aspects of these images. Ultimately, a combination of textual analysis and visual analysis may offer the most robust opportunities for in-depth subject analysis and representation of image collections, though this could represent substantial work.

No matter the route taken, scaling up the procedure described here to other collections would be greatly facilitated by further automation, particularly in the facet analysis and assignment phases. In the facet analysis phase, the objective was to determine which facets were present and what linguistic and semantic patterns those facets might be indicated by. The facets and patterns identified provided opportunities to build a mapping between facets and the keywords in the form of a dictionary. This dictionary may enable the construction of some algorithms to automate the facet analysis of other collections. Doing so would also enrich the current dictionary with additional keywords, though further consideration is needed of how appropriate this dictionary could be for other kinds of collections from other cultural settings.

In the facet assignment phase, the keywords are analyzed to determine which overt and covert facet values are associated with them (e.g., father signifying “father,” “male,” and “adult”). In addition, other terminological sources are consulted for closest matches on these facet values. Though much of this initial work was done manually, automated, linked data approaches hold potentials for scaling up. For example, during the facet assignment phase, mapping could be performed and written into structured statements, as the examples show in Table 7. Envisioning the dictionary as an ontology, facets are classes and keywords or phrases are instances of these classes. Some facets may have sub-facets while others would have only direct instances such as those for the nationality facet. Each facet has an array of keywords that can be used to construct programs for automating the facet assignment.

This ontology could be enhanced by including terms/entities from various controlled vocabularies as further instances of each facet. For example, in LCSH, the term “Dwarfism” (http://id.loc.gov/authorities/subjects/sh85040186) represents a concept that serves as an instance of the condition facet. Fully leveraging this kind of linked data set would require well-designed procedures to build up the relationships between various terminologies and instances. The results, however, could prove useful in applying faceted subject analysis to other collections of materials dealing with sideshows and other aspects of American performance in the nineteenth and twentieth centuries. In addition, this ontology could be used to enhance user search interactions, providing search suggestions and augmenting results with additional terminologies. The overall applicability and potentials of this ontology for a wider array of collections and systems warrant separate investigation.

### 6.0 Conclusion

While any subject representation of historical materials runs the risk of oversimplifying or obscuring original context, the faceted approach pursued in this paper offers one means of moving beyond traditionally reductivist approaches to representing persons, presenting instead a more holistic, multi-faceted view of personal identity. By leveraging existing metadata alongside perspectives from additional terminological sources, our reanalysis of the Becker-Eisenmann collection works to build bridges between historical vocabulary and the language of contemporary communities, and in doing so, more clearly conveys the contexts around these materials. Beyond the identity framework developed in this paper, other contributions include the keyword/facet/value dictionary that could be utilized in the analysis of other archival collections, as well as the beginnings of an automated procedure that can be expanded upon and applied in future research.

As we look toward applying the findings of this study more broadly, several limitations of the present work must be

<table>
<thead>
<tr>
<th>Facet</th>
<th>Relation</th>
<th>Sub-facets</th>
<th>Keywords/Instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>hasGroup</td>
<td>child, adult</td>
<td>{boy, girl, baby, child} {woman, lady, man, couple, wife}</td>
</tr>
<tr>
<td>Nationality</td>
<td>hasIndividual</td>
<td></td>
<td>{Australian, Burmese, Chinese, Circassian, Egyptian, German, Russian}</td>
</tr>
<tr>
<td>Condition</td>
<td>hasGroup</td>
<td>natural, decorative</td>
<td>{dwarf, midget, giant, hairy, dog-faced, albino, fat, legless, armless, hunchbacked} {tattooed}</td>
</tr>
</tbody>
</table>

Table 7. Examples mapping between facets and keywords as instances.
acknowledged. As expressed above, the construction of our faceted framework reflects our own contemporary perspectives on identity; persons from other times or cultures might arrive at a different allotment of identity facets. Similarly, facet analysis and assignment required a certain level of interpretation on the part of the researchers. Further work should acknowledge and attempt to address such limitations. Most critically, gathering user perspectives on these resources and the meaningfulness and usefulness of the metadata generated in this study would allow for a better assessment of the procedures presented here. Other opportunities exist to further explore the generalizability of this work, including the application of this faceted metadata to resource description and discovery in similar archival collections of visual materials. Finally, additional work to further progress the automation of these procedures could allow for more efficient, large-scale implementation of faceted identity metadata.

References


The Challenges of Semantic Interoperability in the Era of eScience on the Web

Maria Luiza de Almeida Campos*, Linair Maria Campos** and Nilson Theobald Barbosa***

Programa de Pós-Graduação em Ciência da Informação, UFF,
Rua Tiradentes 148, Ingá, Niterói, RJ, 24210-510, Brazil,
*<maria.almeida@pq.cnpq.br>, **<lmcampos@id.uff.br>, ***<nilson@tbarbosa.org>

Maria Luiza de Almeida Campos holds a PhD in information science and is Researcher and Professor of the postgraduate programs at Universidade Federal Fluminense, Niterói, Rio de Janeiro, Brazil. Her areas of interest are knowledge organization, models and theories of knowledge representation, terminology, semantic interoperability and foundational ontologies.

Linair Maria Campos is Researcher and Professor in the Department of information science, graduate and postgraduate programs, at Universidade Federal Fluminense. She has a master’s degree in informatics from Universidade Federal do Rio de Janeiro and a DSc in information science from Universidade Federal Fluminense. Research interests include knowledge organization, models and theories of knowledge representation, terminology and foundational ontologies.

Nilson Theobald Barbosa has a master’s degree in information science from Universidade Federal Fluminense and presently is a doctoral student of information science and worked as a substitute professor at Universidade Federal Fluminense. He is also a systems support analyst at Universidade Federal do Rio de Janeiro. Research interests include knowledge organization, models and theories of knowledge representation, semantic web technologies and interoperability of vocabularies.


Abstract: This article is an exploratory study that intends to present semantic interoperability initiatives in the area of information science and computer science. It discusses these initiatives, focusing on open science and eScience, aiming at pointing out similarities and differences in the methodologies used for an intelligent retrieval of data in heterogeneous environments that do not benefit a universal language. We will illustrate seminal initiatives regarding the compatibility of languages in information science and the alignment and semantic mapping in the area of computer science, emphasizing what they have in common and their differences, showing that those seminal initiatives are still relevant nowadays. In this perspective, we want to bring up discussions about methodologies that supports a kind of “intermediate language” that allows the compatibility of research resources, semantically interconnecting their contents without forcing the use of a single language, but of a switching language, such as a metalanguage that guarantees to keep the differences of each community of speech.

Received: 24 August 2019; Revised: 13 December 2019, 28 January 2019, 22 May 2019; Accepted: 10 July 2020

Keywords: data, information, semantic interoperability, compatibility
1.0 Introduction

In this article we discuss the issues concerning semantic interoperability and the methodologies used to enable intelligent data retrieval. In this sense, we investigate methodologies that allow the development of an “intermediate language” that enables the matching of research resources, semantically interconnecting their contents in different contexts, in open data and eScience. The question that arises is how to achieve semantic interoperability between the existing sources of information in these contexts, ensuring their compatibility without forcing the conversion of each one into a single vocabulary. Thus, in this article, we intend to point out similarities and differences in the methodologies investigated and used in the areas of information science and computer science aiming at intelligent retrieval approaches that do not rely on a universal language for data interoperability, showing that seminal methods and strategies proposed by information science should be allied to modern methods proposed by computer science to address the interoperability issue nowadays.

At the end of the nineteenth century, Paul Otlet and Henri La Fontaine conceived an environment where world scientific production could be found based on an expressive effort to collect and organize scientific information. This visionary achievement paved the way for other initiatives that would emerge in the area of knowledge and information organization, with a focus on the document. Otlet also conceived a future world where information in various media would be available to the user at his desk on a screen that would provide not only texts but images, sounds and videos from media such as television, microfilm and cinema (Rayward 1991). After a while, in 1945, Vannevar Bush imagined these information resources connected. So, he was considered the forerunner of hypertext (Robredo 2011). Nowadays, a few decades ahead, we are witnessing these future visions becoming true, and we have been going through an era where the access to information is more widespread and mobile, but, in addition, we live the possibility of access to an immense and diverse volume of data (structured, semi-structured and unstructured), in an open format and often interconnected with one another. We have access to another world of potential information, where a user is inserted whose focus is no longer to find and access documents but to obtain and make sense of this huge volume of data, asking questions for which answers are expected based on the available data. The borders for knowledge acquisition have been stretched out. In addition to it, we are challenged to provide the technological means and to think the requirements that the technologies must reach. We also need to gather differentiated kinds of expertise and efforts in order to ask the pertinent questions, make sense and interpret the answers obtained based on these data. There are many advances in this regard, and in this context DBpedia and Wikidata stand out as central data sources for the embryo of a semantic web being built in a collective and democratic way, which we can already perceive as a vast and varied worldwide database, whose content can be searched by the average user, albeit with some small degree of difficulty (Burgstaller-Muehlbacher 2016).

On the other hand, there are still many challenges to be faced, in particular when taking account the compatibility of vocabularies during the access to this increasingly open mass of data connected in the web. In this context of open data on the web, the relevance of the access to scientific research data stands out as a support for its reuse, to the generation of new knowledge, research visibility and other factors.

The current scenario that is related to the processes, which involve scientific research and the sharing and reuse of information among researchers, refers to the creation of mechanisms in an informational space that makes possible the scientific dissemination of methods, resources and products, the fruits of the activity of research. There is an increasing worldwide interest in an open science at all levels from scientific publications to scientific research data, through all stages and processes involved in scientific research, like public and private funding, resources used, accountability for the society, human resources utilized and similar services. On the other hand, as a result of research activities and the collective willingness to allow and give access to this information, the exponential increase in the amount of information creates a significant difficulty to find relevant information, also strangled by the enormous variety of areas of knowledge and institutions, a large number of data formats and various metadata used. Thus, we consider that in the field of information science it is necessary to investigate mechanisms, which make it possible to create a network of information that, despite the diversity and heterogeneity, can link information spread across several institutions and different areas of knowledge and make sense of them. In this context, ontologies are being used in online scientific activities, or eScience, mainly in papers related to the management and integration of data resources and workflows, with the view to allow a more explicit representation of scientific artifacts (Brodaric and Gahegan 2010). It is in this context, that we present the interoperability issue as a condition for heterogeneous information systems to interconnect semantically similar information.

For Payette et al. (1999), interoperability is defined as the ability of the components or services to be functionally and logically interchangeable, because they have been implemented according to a set of well-defined and known interfaces. However, the access to heterogeneous systems can involve documents (or datasets) from different research groups and institutions and even different countries, that
have their information indexed quite differently, whether linguistically, structurally or in distinct conceptual schemas. In a diversified environment in some or all these aspects, it may not be possible to use the same interfaces with nomenclatures, and retrieval processes may lead to unsatisfactory results. In such cases, the users may be led to do multiple checking to get relevant results or sometimes even get some results. For the establishment of interoperability between different systems or indexing languages, some methods, at different levels, can be adopted.

The substantial increase in technologies that promote open interfaces and the creation of comprehensive metadata standards, which can represent academic research environments, are undoubtedly great steps towards advancing the process of interoperability. The problem is in the treatment of this research data at a level not only descriptive but also semantic. This includes addressing related issues and discussing the possibilities of sharing such data by relating their content. Within the semantic web environment, the study and the creation of ontologies have relevant aspect in the solution of the problem. Ontologies are a powerful way of interrelating systems. They are elaborated, mostly, aiming the structuring of knowledge bases or to be used as semantic tools in the support to the interoperability between information systems (Campos 2007).

In this case, it is very important for institutions from any field, to create their controlled and specialized ontologies or vocabularies, but the enormous semantic heterogeneity is not capable of allowing semantic interoperability by itself, since it would be unlikely to create a global ontology to which everyone would simply adopt. Therefore, in the real world it is necessary to create mechanisms that enable this semantic interoperability between heterogeneous systems. Interoperation between two systems seems to be possible only when we can overcome the linguistic, terminological, and verbal differences that are defined by the idiosyncratic conditions of each environment. So, we can compare and link the definitions and concepts associated with each object. In this sense, we are engaged to investigate methodologies, both in the scope of information science and computer science, which permit us to develop a kind of “intermediate language,” that allows the compatibility of research resources, semantically interconnecting their contents.

Dahlberg (1981), in the area of information science, proposes the construction of a conceptual compatibility matrix based on her analytical-synthetic method. The conceptual compatibility matrix is a mapping of the semantic potentiality of the languages studied, providing the results of the compatibility analysis between languages from the semantic and structural point of view. Another work published in the area of information science that deals with the compatibility theme is the thesaurus reconciliation method proposed by Neville (1970), which is based on the principle that compatibilization must consider concepts (the conceptual contents of the descriptors, which are expressed in the definitions) and not the descriptors only. This method proposes an intermediate language approach, based on the numerical coding of concepts through which it is possible to establish the conceptual equivalence of descriptors of different languages, considering also the one to n correspondence treatment between the terms to be matched.

In the field of computer science, related to the studies of semantic representation languages, the literature has presented several forms of compatibilization. Here we emphasize those related to the concept of alignment (Bruijn et al. 2006), which allows the creation of links between ontologies, while preserving them without change.

This paper intends to contribute with research addressed to discussing and finding solutions to the problems of semantic opacity already pointed out by Pierre Lévy in his works, when he affirms that we are in the process of establishing a common participatory memory to all humankind. For Lévy (2014), the limitation we have today, at the beginning of the twenty-first century, to explore this immense memory of data are the problems of understanding its terminological meaning, incompatibility of classification systems and linguistic and cultural diversity. Under these circumstances, the lack of models that can be computationally treatable prevents the automation of most of the cognitive operations of analysis, selection, synthesis and interconnection of potential information, and, thus “we do not know yet how to systematically transform this ocean of data into knowledge, and still less how to transform the digital media into a reflective observatory of our collective intelligences” (Lévy 2014, 23).

The statement of Lévy, mentioned above, motivated us to take this path, that is, to raise initiatives that can provide some solutions for an intelligent retrieval for the immensity of data produced today. Then, we are starting these studies in our research group. Hence, this article has an exploratory objective, of raising questions that go through this issue. Thereafter, this article is then organized, besides this introduction, to present the questions that involve the great diversity of data currently produced, and the problem of intelligent access to this data. Section 2.0 “Research data in a new context: open science and eScience,” presents specific questions that involve the research data in the context of open science and eScience, in addition to raising a current discussion, that is, restricted or unrestricted access to the retrieval of such data. It also raises the matter of the possibilities of relating this data through its contents in a significant way in an environment, which semantic technologies are already taking their first steps. Thus, in Section 3.0 “Semantic compatibility studies: contributions of information science and computer science,” we present initiatives in information science and computer science of semantic compati-
bility and interoperability showing similarities and different methodologies presented in prospects of mapping possibilities for an intelligent retrieval to environments where the research presents itself. And finally, we present in our final considerations aspects that we consider fundamental for such initiatives to be undertaken.

2.0 Research data in a new context: open science and eScience

Research data as stated by the Organization for Economic Cooperation and Development (Pilat and Fukasaku 2007) are records of facts used as primary sources in scientific research and are generally accepted in the scientific community as necessary for the validation of the research results. These data, according to the nature of the knowledge in which they are generated, can be identified as numbers, images, software, algorithms, animations, texts, and other resembled items. The literature presents a range of typologies about the classification of research data (Borgman 2010; Harvey 2010; Lyon 2007; NSF 2007) where each classificatory form is linked to certain purposes. Then, as we can observe, the research data are collected to serve several purposes, using various methods, being possible either the purposes or the methods to influence on the data classification; moreover, these classifications also relate to how these researchers are willing to share their data with their peers (Sales 2014).

Research data are also called scientific data and have always been part of the process of doing scientific research, but they were not considered and are not always presented in the products/research results. We can discriminate such data as information objects produced, collected or used throughout scientific research. They are distinct and have a heterogeneous, diverse and complex nature. With the advent of the eScience and the open science movement, scientific data have gained prominence and recognition of its importance not only to the scientific community but also to society. These two concepts have a certain conceptual correlation, although they are not similar, as we can observe.

The term eScience was created by John Taylor in 2001 as a new way of doing science. That is, by establishing a global collaboration in key areas of science and every generation of infrastructure that will enable the sharing of information among researchers. From Taylor, other researchers have contributed to the ideals of eScience (Gray et al. 2005; Hey and Hey 2006; Hey, Tansley and Tolle 2009; Taylor 2001) in which ICTs and data repositories should play a prominent role. The creation of such digital repositories of scientific data distributed in various spaces where the research that is being done will require management services similar to the ones of the conventional digital libraries, as well as other data management services. In this case, it is important to point out that the availability of this data is expected for the relation between specialists. To make science (or e-science as it can be said today) requires manipulating and storing large quantities of these data, and still requires that such data may be available, to be shared and reused by other researchers and research groups, with similar objectives or not from the group that originally used or produced them. Though, disclosure in eScience, in a way, is more used to the availability of scientific data among peers.

In relation to the open science, its primary concern is to make research activity more transparent, more collaborative, and more efficient, in which scientific knowledge must be understood as world heritage. According to Akmon et al. (2011), scientists can no longer assume that the data generated during their research has value only to themselves or to their research group. Instead, they need to anticipate future uses of the data and, thus, preserve them and provide descriptions that facilitate their reuse. But these scientists do not have the knowledge, the techniques, the interest or the time to do it.

It is in this context that the proposals for the creation of digital curators of scientific research data, in accordance with Lee and Tibbo (2007), have as main tasks to boost the availability of data for the scientific community, to offer effective storage, to transform data, to preserve for a long term period, to make available authentic digital data to be reproduced and reused, to develop reliable digital repositories, to generate principles of metadata creation and capture and to use open standards for file formats and conversions. Besides, while data is central to the scientific process, metadata is central to the operation of digital files and to the curation of digital data, but the essence of the problem is to ensure that all the metadata and their associated documentation, that are required for the efficient use of data are correctly generated, understood and effectively accessible to its legitimate users (Davenhall 2011).

These repositories of scientific research data, despite being a major advance in the description, management and sharing of research data, confront a major problem. The extreme complexity of the data models involved and the enormous diversity of objects and domains to be represented form an almost insurmountable barrier to the efficient management of these environments, preventing its main purpose, which is, sharing, discovery and reuse of these data to be performed properly. It is a clear example that only the computer and computing infrastructure that forms the basis of the global information network are not enough to solve the existing problem.

The issue we want to address here is what contribution information science can make to open science and eScience. We discussed and used research data and eScience as a context for our argument, because despite the importance and all the effort dedicated to the preservation, dissemination
and sharing of this data, there is still a problem to be solved nowadays, not only eScience data, but digital data in general; interoperability between heterogeneous databases, especially their semantic interoperability, so that researchers can reuse research data without the need to modify the vocabularies that describe them, that is, maintaining the local identities and idiosyncrasies that define the repositories available on the web.

Although the debate on restricted or unrestricted access is at stake when we discuss subjects related to scientific data from which we should not be absent. Here we are especially interested in discussing the modus operandi of how this access can be realized and investigating the possibilities of relating these data through its contents, in a significant way in an environment in which the semantic technologies are already giving their first steps. Therefore, it is our interest to join efforts and go through one of the aspects that will enable such access, whether restricted or unrestricted, that data can be gathered by their conceptual similarities. In this approach, this study aims to investigate issues involving semantic compatibility and semantic interoperability, which are fundamental for intelligent retrieval not only in research environments but in the whole web.

3.0 The studies of semantic interoperability: contributions of information science and computer science

For W3C, semantic interoperability means enabling different agents, services and applications to exchange information, data, in addition to knowledge, notably, both on the web and beyond it. For the purpose of making this type of interoperability possible, the agents involved need to share the same vocabulary comprehended by all or to create correspondences or mapping between different vocabularies (Ushold and Menzel 2005).

According to Degoulet, Fieschi and Attali (1997, vol. 9, 5), in order to achieve semantic interoperability, it is necessary the compatibilization of these vocabularies from the processes of creating and interpreting messages, which depends mainly on three key factors:

- Terminology or nomenclature
  It is the way in which concepts are expressed in a controlled vocabulary having the referential status of a field of knowledge or activity.
- Explicit semantic relationships related to terminology
  Relationships can be organized in a semantic network. This network aims to structure the connections between the concepts of the domain, specifying their nature; synonymy connections, heteronymy, derivation, etc. (for example, SBP and systolic blood pressure or angina and angina pectoris for synonymy connections).
- Representation of concepts in messages
  In general, two different institutions model and represent concepts in different ways. Simple examples are age (modeling created based on the age or from the date of birth ...) and the address of a certain person (separation of street number, ZIP code, number of lines ...). We could cite innumerable cases, as the modeling of the existence of a symptom by stage (or severity) or simply by yes or no.

The search for interoperability between data and records extends across all domains, along with areas of knowledge and at all levels. It is evident, as an example among many others, its need in healthcare, where patients, diseases and drug data need to be exchanged between different institutions, especially using as an example, one of the points where the problem is reflected: the patient’s chart. Although the institutions and professionals of ICT and healthcare of these institutions do not know how to conceptualize correctly what are interoperability standards or even ignore this concept, according to the work of Miranda and Pinto (2015).

Regardless of global efforts, whether governmental or from the research institutes themselves, to promote open science with free access to publications and scientific research data, as shown above, this attempt runs up against the difficulty of interoperating these data between repositories and between different fields of research and human knowledge. The formation of transdisciplinary knowledge networks, that will allow a leap forward in the development of sciences, scientific research and innovation, depends on the ability of researchers, through their publications and research data, to interoperate and intelligently and semantically share their research and results.

As Lévy (2014, 231) states, “so much information is available in the digital world that the obstacle to accessing it is due to this abundance itself: how to find the relevant information needle in the giant haystack of digital data?” According to the author, in order for this information economy not to lose the value of its main good, it must become a semantic information economy, since the value of information depends on its meaning. That is, the information economy should be able to modify the meaningful contexts, the practical environments where the meaning is. In other words, the information economy should be able to model the meaningful contexts, the practical environments where meaning is. Since information availability is no longer a problem to be overcome, informational goods gain and lose value as a function of their significance and relevance to the communities that use them. Then, it is necessary to imagine and build sociotechnical devices that are cable of answering
a central question for a given user: how to find the information that has the most value.

Thereby, by proposing the bases of the semantic web, Tim Berners-Lee and others were saying that the original web operation, which was created and developed by Berners-Lee himself, could no longer meet the wishes of its users (Berners-Lee et al. 2001). The page-based web whether static, as it is at the beginning or dynamic, could not sort out the new information retrieval issues that are demanded by users. The current web was designed to be used by humans and we need to evolve into a web used by intelligent software agents and intelligent devices. These bases and semantic guidelines for this new web paradigm includes the capacity to express meaning, to represent knowledge, to use ontologies and to have software agents as users.

Dealing with interoperability with the intention to disseminate, share and reuse data, records and information is not a new problem to be faced by humanity and the scientists. On the contrary, the need to share information from different sources, with form and meaning, has already been in the information science for some time. On the way to creating a useful memory, the Belgian Paul Otlet, since before the Second World War, had proposed the creation of a universal library, its indexation and also discussed its theoretical problems. When Otlet created the Universal Decimal Classification (UDC), he transformed the classification developed by Dewey into a faceted language and, in parallel, popularized in Europe the use of microfiche, which had already been used in the United States of America. In 1930, Otlet showed that the documentary universe was growing strongly and predicted that electronic technologies would make information ubiquitous. His unfinished project of constructing a collective and universal memory for mankind can be seen in his 1934 work, called The Treaty of Documentation, where the author proposes an intellectual network universe, accessed by a universal classification system, which would be mandatorily updated constantly by the links created by the users between the various documents. In this case, even before the innovative and revolutionary constructions proposed by Vannevar Bush, Douglas Engelbart and Ted Nelson, Otlet presented the project to formulate and build the principle of hypertextual interconnection (Lévy 2014).

In this quest for interoperability, knowledge organization systems are essential to represent the different environments of organizations, institutions and research groups that seek to disseminate and share their information. These KOS can vary greatly in different aspects, i.e., knowledge domain, language, granularity or structure, making it difficult to recover information between different vocabularies. The definitions presented by ISO 25964—Thesauri and Interoperability with other Vocabularies (ISO 2011; ISO 2013)—can help solve these problems. It covers the interoperability of thesaurus, taxonomies, ontologies, classification schemes and other KOS. Its norms and definitions focus mainly on the implementation of mappings and address the characteristics of the different KOS vocabularies that make these mappings possible. However, despite the importance of the existence and formulation of these standards, it is not possible to guarantee that different vocabularies spread around the world today on the web can interoperate immediately, hence the importance of studying and improving the compatibility processes proposed by CI and by CC, which we will see in the next sections of this article.

Thereby, there are also the issues addressed by information science, in order to solve the problem of intelligent information retrieval adding semantics to these searches through methodological procedures that do not focus only on a universal language, but that allows a metalanguage in which the “languages” of the various contexts could be respected. This is a problem to be solved nowadays. In other words, how to make possible an intelligent retrieval of information with such diversity and volume of data generation. This diversity comes from the space composed of different areas of knowledge, ontologies, vocabularies, languages and cultures. Even in controlled environments, like a single company, for example, we have difficulties in establishing or even creating a single vocabulary, that caters to all producers and consumers of information. Thus, it is our understanding that intelligent information retrieval occurs when it is possible to retrieve information from different bases and indexed by heterogeneous vocabularies without the need to intervene or modify the original bases.

In this sense, we will discuss initiatives of language compatibility in information science and alignment and semantic mapping in computer science, which introduce the notion of a metalanguage. This metalanguage must be formed as an intermediary language between the different source vocabularies allowing different actors to navigate through this language and, in a contextual and semantic way, to retrieve the desired information, no longer based on character strings, nevertheless, rather in their meaning, as we will see later.

3.1 The studies of semantic compatibility in information science

When debating about interoperability it is important to declare the high importance, recognized for a long time, of the study of “vocabularies” or “languages” in the organization of knowledge and information retrieval, within the field of library and information science, as pointed out by Lancaster (1986) and Buckland (1999), among others. Recent and growing studies in the field of semantic web, ontology engineering, metadata and open linked data and markup ontology languages that spread across several disciplines and sci-
scientific fields, clearly show the critical role played by vocabularies to represent, access and retrieve information and knowledge (Park 2006; Hovy et al. 2001).

The studies of compatibility and convertibility between languages in the field of information science, more specifically in those concerning documentary languages, from the 1960s of the last century, intended to create tools that would allow the conversion of different languages, in order to enable the user access to different data sources. It is considered that in the 1960s these studies stood out, due to the “information explosion” and the subsequent loss of information control, because of the proliferation of data sources. At the same time, the possibility of different US agencies accessing each other’s content and a possible speeding of service delivery encouraged the studies about compatibility (Lancaster 1986). Hence, the research in the area was contextualized at a political moment when information reached strategic and decisive status. In the 1970s, there was a decline in the studies about compatibility, due to the dissemination of research automation. Then, in the 1980s, the studies about compatibility arose again, now applied to the computational environment and to automated language issues (semantic and syntactic problems). So, in the 1990s, these studies dealt with the integration between languages of computational systems with emphasis on the elaboration of ontologies, that are used as semantic tools for the purpose of allowing interoperability between systems (Campos 2010; Souza and Campos 2007).

Compatibility can be understood in two aspects: semantic and structural compatibility. Semantic compatibility is the ability of two vocabularies to have similarity between the conceptual contents of their terms while structural compatibility can be understood as linguistic compatibility (Glushkov et al. 1978). In either of these two aspects, we emphasize that in information science the search for compatibility of information systems and knowledge organizational systems is not a minor problem and is strongly related to one of its essential objectives, which is to allow and simplify the connections between those who need sources of information on a given subject and the potential relevant documents to meet this demand. In order to be able to advance in the solution of this problem, especially in the current times of massive production of information sources and advancement of information technologies, it is necessary to understand that the compatibility of languages and vocabularies is not an intrinsic quality of these systems, but rather, it is a goal to be achieved (Maniez 1997), as discussed below.

Although the authors debate about compatibility in thesauri, this discussion of semantic and linguistic aspects in the process of compatibility can be applied to ontologies. From the methods of compatibility and conversion of languages based on the integration of vocabularies, two stand out eminently. They are the “thesaurus reconciliation” method, suggested by Neville (1970, 1972) and the “conceptual compatibility matrix,” proposed by Dahlberg (1981b, 1983). Neville’s method is based on the principle that concepts must be made compatible, (the conceptual contents of the descriptors, which are expressed by the definitions) not the descriptors alone. This method offers an intermediate language approach founded on the numerical coding of concepts through which it becomes possible to establish the conceptual equivalence of descriptors of different languages. It also considers that, within a common thematic area, vocabularies should cover the same concepts even though there may be different terms for naming the same concept among these different vocabularies. Based on this principle, their strategy is established on identifying similar concepts and encoding them uniquely in each vocabulary. This coding would then allow the keywords of a vocabulary to be mapped to other vocabularies of the same subject that shared this coding scheme. For this, Neville (1972) recommended an intermediate language way, which implements this numerical coding of concepts making possible to establish the conceptual equivalence of descriptors of different languages, denominated by the author as reconciliation, that is, the possibility of integration and approximation of systems, which contemplate the same type of literature, but adopt different thesauri (Neville 1972, 622). In the author’s plan, each term in each participating thesaurus receives a code with the inclusion of remissive to the key-words. Moreover, a key to the encoding is given, so that your application in third-party keywords will generate the corresponding keyword in the source thesaurus. This key is different for each thesaurus; it functions as a conversion mechanism.

The establishment of correspondences between concepts does not necessarily imply a one-to-one correspondence. There may be cases, for example, where a more specific concept in one of the vocabularies is covered by a broader concept in the other vocabulary, or there may be cases where there is no correspondence in the other vocabulary for a particular concept of the vocabulary of origin. Neville (1970) embraces a thesaurus as a basis and follows eleven levels of matching between the terms of the two vocabularies analyzing them from the exact match between terms, use of synonyms and homonyms, among other things. Like Neville, Dahlberg (1981b) also introduces a mapping of the semantic and verbal potentialities between languages, which she calls “compatibility matrix.”

Dahlberg (1981b) states that the most sophisticated use of a compatibility matrix would be if it worked as a black box, in which one would enter a descriptor in an indexing language (IL), and it would be possible to obtain the corresponding descriptor in the other ILs involved. Dahlberg’s conceptual compatibility matrix (1981b) is a mapping of the semantic potentiality of the languages studied, provid-
ing the results of the analysis of the compatibility between languages through the semantic and structural perspectives. The compatibility between languages, as said by the author, comprises three phases: 1) conceptual coincidence—when two concepts combine their characteristics—degree of equivalence; 2) conceptual correspondence—two concepts combine most of their characteristics—similarity; and, 3) conceptual correlation—two concepts are correlated through mathematical symbols establishing a correlation measure when they have different levels of detail, or when the relation between them is not of similarity (Dahlberg 1981b).

While Neville (1972) propounds the creation of an intermediate language in which the potentialities of reconciliation between two languages are presented, Dahlberg (1981b) goes a little further, since, besides presenting principles for the mapping, she also presents a proposal, which introduces the concept of “compatibility rate” in which one can quantify the degree of compatibility between the languages under analysis. Additionally, Dahlberg (1981b) uses the term “ordered systems” to name the various “terminological tools” that should be compatible and conceptualizes them as any instrument used in the organization, description and knowledge retrieval, composed of verbal or notational expressions for concepts and their relations, arranged in an ordered way. Dahlberg cites as examples classification schemes, thesauri, subject headings or another identical instruments.

Thus, when adopting “ordered systems,” it extends not only the concept but also the universe of application of the compatibility methodology, since it makes the definition of the instrument to be compatibilized flexible and can include the questions asked by the users analyzed within a context that presents an internal organization offered by the retrieval system. Another issue presented in this study is the role of the definition in the mapping of the semantic potentialities of compatibility, when the established correlation between ordered systems should not be performed only at the level of the terms and their descriptors. The terms are only the bearers of the information and can only display the information about the content of the represented concept if the necessary relationships have been established through a definition.

A fundamental point presented by Dahlberg (1981b) in her methodology is the need that for each term there is a kind of annotation about how the term is structured in a given ordered system; she called the this information about the term “concept record.” Therefore, for two languages to be compatible, it is necessary that each establishes a register of concepts. According to Dahlberg (1983, 6):

In establishing an ordering system one must attempt to accumulate the necessary knowledge about the ref-
erents by a “concept record” ... If an ordering system has not been developed with the help of concept records it is necessary to establish such records at the time when the said comparisons with other OSs should be made of each single class and concept on every level of the hierarchies foreseen in order that the analyzed conceptual data of the concepts in question can be introduced into the comparisons.

The register of concepts, as told by to the author (Dahlberg 1983, 6), consists of a series of necessary fields:

- 00 – running number;
- 01 – name of concept or class of concepts;
- 02 – notation;
- 03 – definition with indication of source;
- 04 – next broader concept (generic and partitive relationship);
- 05 – highest concept in hierarchy;
- 06 – subject(s) field of a concept;
- 07 – other names of a concept or class of concepts (synonyms);
- 08 – source of concept; code for ordering system and
- 09 – remarks concerning corresponding concepts in other ordering systems.

Furthermore, the author suggests possible alternatives to these necessary fields: i) the name of the concept in other languages; ii) the category of the form of the concept (form category), which indicates if this is an object, a process or a quality; iii) additional information about the concept; and iv) related concepts.

The concept record supports the conceptual comparison between ordered systems, which is generally founded on the analysis of the hierarchical structure in which the term is situated, in addition to its definition and possible comments, to conclude whether the terms are or are not corresponding to the same concept.

For the preparation of a compatibility matrix, the first step is the verbal or linguistic joint of the terms, which can be automated and recorded in a preliminary matrix. From the analysis of the percentage of terms that were possible to match, we evaluate the feasibility of continuing with the compatibilization of vocabularies. The linguistic compatibility, however, does not ensure that the coincidences found are actually a conceptual correspondence, due, for example, to the possibility of homonymy. In addition, if different languages use different nomenclature for terms with the same meaning or present concepts at different levels of detail, the linguistic match is not enough to detect such occurrences.

In this way, the preliminary matrix obtained must be complemented, in a second stage, by means of the analysis of the concepts so that a semantic correspondence can be
established. It is at this moment that the concept register is established to be able to infer the level of conceptual compatibility, that is, conceptual coincidence, conceptual correspondence and conceptual correlation, as presented above. As a result of this complementary conceptual analysis, we obtain the final compatibility matrix, which establishes, additionally to the correspondence of the concepts, a measure of compatibility and the correspondence type (“<”, “>”, etc.) as it was mentioned previously.

Both Neville’s (1970, 1972) and Dahlberg’s (1981a, 1981b, 1983) studies, although stemming from thesaurus activities, are considered to be appropriate for the investigation of principles related to the semantic aspects of terminological compatibility in heterogeneous environments of scientific data, in addition to the fact that they present elements to discuss the level of conceptual similarity independent of the types and data formats involved.

Nowadays, a significant number of conceptual and structural methodologies for vocabulary creation resemble those found in the history of pre-web knowledge organization systems. What makes them different are the approaches supported by semantic technologies, in consonance with LOD principles. With the growing number of KOSs being published in standardized and machine-understandable formats, institutions can refine and augment their data from external sources, accomplishing a major achievement of reusability of these vocabularies (Zeng and Mayr 2019). But while we have a steadily increasing advancement in vocabulary-building and sharing technologies, especially semantic web-related technologies for automated processes, on the other hand, we still need to make great progress in the methods and strategies for applying these technologies. Therefore, to advance and effectively reach the possibility of working in an open environment aiming at the interconnection of heterogeneous data with different vocabularies, without having to change or intervene in these vocabularies, we highlight two aspects of great value: to rescue the previous efforts mentioned above concerning the classical methods of information science, and to establish its similarities and differences from modern approaches, especially those directed to automatic processes.

Even with the incredible advance of information technologies, we have several mapping processes done at a syntactic level, that is, words, phrases and context, instead of being done at a semantic level. Advances obtained by the new processes of artificial intelligence and machine learning have great potential in reducing conflicts and promoting interoperability, in particular semantic interoperability (Zeng 2019), but despite these advances, the mapping of concepts in a semantic manner is still a challenge for those who depend on automatic mapping.

As we can see from methods and processes covered in this section, our view is that seminal information science studies for vocabulary matching, even if they were created in a pre-web era, are essential to moving forward in solving the problem of web interoperability and in creating automatic compatibility processes.

### 3.2 The semantic alignment studies in computer science

Leiva-Mederos et al. (2017) emphasize that there are several mechanisms to tackle semantic interoperability and that one of them is semantic alignment. The authors believe that semantic alignment fundamentally means finding the correspondence between distinct vocabularies. Li, Yang and Liu (2008) show that semantic alignment is the sum of a variety of methods and can be applied to many different types of thematic domains.

In the field of computer science, related to the studies of languages for semantic representation, by the end of the 1990s, ontologies began to appear. Ontologies, as a terminological artifact, can be compatibilized in different ways, depending on the needs that must be met and their availability. In this work, aiming the comparison with the initiatives of information science, for the purpose of intelligent data retrieval, we will treat ontologies as terminologies with the goal of representation and information retrieval.

In this field, the literature has presented several forms of compatibilization (Bruijn et al. 2006; Pinto and Martins 2001; Ziegler and Dittrich 2004, vol. 12). In this paper we will investigate the concepts of alignment (Bruijn et al. 2006; Choi, Song and Han 2006), which allow us to generate a set of links between ontologies. The alignment process, unlike others (e.g., join, integration) in relation to its result, instead of generating an additional ontology, keeps the reused ontologies unchanged and in their source locations but generates a set of (links) between these ontologies. These links contain a set of information about how to make reused ontologies compatible and they are expressed in a separate (physically existing) persistent model.

Alignments can be used for various purposes, such as the transformation of one ontology into another, or to allow queries or searches on features described with ontologies (Euzenat et al. 2004). Thus, a search can be made, for example, with terms of an ontology (O1) over features described with a another ontology (O2), because the alignment allows the automatic matching of the questions to the ontology model mentioned.

The set of links expressed in a persistent model produced by the alignment process is a mapping between the ontologies. There are different definitions in the literature for mapping (Ding and Foo 2002). In the context of this work, we adopt mapping as a formal expression, that establishes a correlation between two elements of distinct ontologies and that is stored in a persistent model, which can be separated...
or incorporated into an ontology. The elements related by the mapping can be the classes (used in the context of this work as a synonym of term), the properties or the attributes contained in the ontologies.

The information contained in the mapping will depend on the type of semantic link found between the elements and the type of formalism used in the ontology to represent its semantics. For example, two elements may be similar (to different degrees), or one may be part of the other or they may have some other type of relationship, which is identified with the help of a domain expert.

In relation to formalism, mappings can also be used to: a) make different formats compatible in the representation of ontology class attributes (for example, an attribute size measured in centimeters or millimeters); b) overcome problems of expressiveness derived from the use of different languages; c) make the shape of a class compatible, for instance, the wing of an airplane can be modeled as a part of the airplane (in this case the wing itself is a class); or, d) as an attribute of the airplane class. Other aspects can also be considered in regards to mapping such as the cardinality of relationships established when mapping, which can be one to one, one to many or many to many (Vola 2008).

Mappings of similarity may express different degrees of similarity. These can be represented as a numerical attribute of the mapping (Felicíssimo and Breitman 2004; Kalfoglou and Schorlemmer 2003) or also by means of a relation, that indicates the type of similarity, such as “equivalent,” “narrower than,” “broader than” (Aleksovski, Kate and Harmelen 2006; Su 2004). In order to determine the degree of similarity, several factors are generally considered: linguistic similarity between terms, compatibility of attributes, positioning of the term in the hierarchical structure of the ontology, among others.

As for the strategy of mapping, there are three ways to implement it: i) between local ontologies; ii) between a global ontology and other local ontologies; or, iii) in an ontology that is the result of the joining of others (Choi, Song and Han 2006). A local ontology reflects the conceptualizations of a domain that are relevant to a particular community, thus mirroring the perspective under which that community represents its world (or part of it) according to its objectives (Bouquet et al. 2002). A global ontology is a combination of several local ontologies in a unified ontology, which in this way contains all the concepts of local ontologies.

In the first way to implement a mapping (i), there are two local ontologies mapped through a separate persistent model (Choi, Song and Han 2006). In the second form (ii), there is a global ontology and one or more local ontologies and a separate persistent model that maps the terms of the global ontology to the terms of the local ontologies. In the third form (iii), there is an ontology, which is the result of a process of joining others and a set of mappings that is stored within the ontology itself, the result of the junction process, in the form of axioms and which maps the terms of the ontologies imported by this final ontology.

Several ontology matching systems have been designed recently and most of them use element-level techniques that aim at lexical information as essential elements, but with only this simple string comparison approach is impracticable to get useful results. On the other hand, there are structure-level techniques that rely on the analysis of the neighborhood of two entities to determine their similarity. But both techniques present weakness as they must extract semantics from the lexical information of entities. A possible solution to address this problem is to rely on external sources like WordNet or Wikipedia in order to obtain semantic similarities among elements. The limitation of this approach is that many vocabularies cannot find their correspondences in WordNet or Wikipedia. Another possible solution is the use the technique of word embeddings similarity, that can represent words as vectors in a semantic space. Even if different methods and strategies have been used, in almost all cases, the intervention of a domain expert in the ontology alignment process is necessary to avoid inconsistencies, showing the limitations of the processes (Zhang et al. 2014; Dhouib et al. 2019; Ardjani et al. 2015).

3.3 Similarities and differences between the two approaches

Based on the exposed approaches and techniques, it is possible to observe some common aspects in these diverse views. Primarily, in the proposals resulting from information science, we can observe a strong concern with the compatibility of the concepts involved and their conceptual contents, not only with the character strings that represent the terms, in other words, their descriptors, since this is a determinant characteristic of this approach. When Dahlberg proposed the semantic mapping, through semantic potentiality, she went beyond the coincidence between the concepts, analyzing its correspondences and correlations and suggesting the creation of a conceptual compatibility matrix, as exposed above. Neville, on the other hand, considers this semantic view through the integration of different thesaurus systems with the creation of an intermediary language, making a conceptual equivalence of the descriptors, not necessarily from one to one.

These proposals, in spite of privileging a semantic and conceptual view, have as an important aspect the fact that they require a great intellectual and manual effort to be implemented, requiring the work of professionals who have the knowledge of the processes to be carried out, and of professional experts in the vocabulary domains to be aligned. With the exponential growth of data—in the internet post-
creation period and especially in the web post-creation period—allied to the growth of the diversity of vocabularies and indexing systems spread all over the world, in different domains, different cultures and different languages, this intellectual and manual work could not, by itself, be capable of meeting in its completeness the tasks of integration and compatibility of the heterogeneous data.

Concurrently, different techniques are developed, mainly by studies linked and supported by computer science so that, supported through computer programs they can automate and move towards integrating heterogeneous data sources. Therefore, we need to guarantee that in this case, as well as almost everything in this interconnected world, the practical application of interdisciplinarity, essential in the eScience era, be possible. The use of working tools for large volumes of data and large vocabularies should use compatibilization procedures which consider conceptual concepts that deal with the semantic similarity between terms.

DeRidder (2007) highlights that the semantic heterogeneity is inevitable, especially when information systems grow in a decentralized way. We can notice that this is what happens with the web. The author, in the context of ontology interoperability, explains that in order to achieve interoperability, there are three basic approaches that can be chosen: a global ontology for which local ontologies are mapped, local mapping between ontologies’ pairs when necessary and a combination of both strategies. For DeRidder, although a single global and heavyweight ontology is a preferential option from the computational point of view, on the other hand, obtaining a global agreement for this ontology seems to be unfeasible. He proposes that an approach based on generality layers, with a group of ontologies for this purpose, instead of only one, would be more likely to succeed.

Park (2006, 30) shows in her study results that point out the need for mediation mechanisms, that it is necessary in order to provide contextual relations between metadata elements and their definitions to ease the mapping processes that can reduce semantic ambiguity. She proposes that the use of concept networks may serve this purpose; however, she suggests that the development of this kind of solution requires additional studies concerning not only the metadata but also the mapping practices. In this context, she highlights that “problems in metadata mapping result from the absence of the context in which a metadata element name and its usage (i.e., definition) occur.”

On the other hand, although human mediation and intervention are still necessary, Greenberg, Spurgin and Crystal (2006) point out that it is not possible to rely exclusively on traditional manually-generated metadata approaches, especially considering the huge number and volume of digital resources in need of these metadata. In a study related to this discussion, focused on metadata generation and that contributes to the understanding of automated processes in this area, Greenberg (2004) positively concludes that there is a potential of these automatic processes for metadata creation.

For Martínez-Ávila et al. (2018), the semantic web, by proposing the creation of machine-readable data by different communities, has led to isolated information systems that deal with their own knowledge and are specific to their domain, limiting potential interoperability since publishing data as linked open data (LOD) is a big step towards making data available and interconnected but not enough to achieve full interoperability. The technology to link this data exists, but performing the linking process requires expert knowledge and does not happen completely automatically. Moving to a LOD environment requires extended interoperable vocabularies for better organization of large data, that is, we need producers and users who are able to provide and use richer semantics and structure in the vocabularies that are used to consume, describe and publish data. This situation could benefit from a greater application of knowledge organization principles (Martínez-Ávila et al. 2018).

According to Chan and Zeng (2002), as the advanced computerized processes were evolving, to achieve or improve this interoperability among vocabularies, computer technologies begun to be used to benefit from this interconnected environment. Thus, some new methods and some conventional ones have come to form a list of widely accepted methods.

In the method called “derivation” or “modeling,” a simpler or specialized vocabulary is developed from a more comprehensive one chosen as a model. In “translation” or “adaptation,” a vocabulary is controlled and developed from terms translated from another language. In the intellectual process of mapping a system, it is developed from the establishment of the equivalence of terms between different controlled vocabularies, requiring a great intellectual effort. This method can be migrated to use the computer technology in its implementation (Chan and Zeng 2002). In “alignment,” for each entity (which can be a concept, an instance or even a relation) in the first vocabulary, we try to find a correspondent entity in a second vocabulary with the same intended meaning in the second vocabulary (Ehriog 2007).

In this environment where computer and information technologies are used, vocabulary alignment and mapping emerge as the main methods to be used in the vocabulary compatibilization process, because they can be applied to a variety of usages and scenarios such as vocabularies and ontologies importing, data schemes integration and links of different versions of ontologies. Thus, the ontologies alignment is a crucial condition nowadays to make possible the interoperability between semantic systems when it allows that the individual relations between elements of multiple
ontologies to be identified (Ehrig 2007; Euzenat and Shvaiko 2007).

It is possible to observe in the scientific publishing of information science and computer science areas, different meanings and conceptualizations for the term’s alignment and mapping (Ding and Foo 2002; Su and Gulla 2006; Ehrig 2007; Fielding et al. 2004). For some authors, such as Ehrig (2007), the alignment is a preliminary stage where it is possible to detect at which point the ontologies overlap each other and where they can be connected to each other. For this compatibility type, the terms to be aligned have the same meaning, but the process foresees that different relations be used and not only equivalence. In this alignment point of view, different relations such as identity, subsumption, instantiation and orthogonality can be used in the links between the terms (Ehrig 2007). For Euzenat and Shvaiko (2007), alignment is based on a term’s matching process where it is sought to identify those terms that express similar concepts, being possible to associate values to express their degree of reliability and similarity (Ehrig 2007).

Regarding mapping, some authors, such as Păslaru-Bontas (2007), consider it as a synonym to alignment, claiming that both terms are used in an interchangeable way in literature and define it as a process of relationship creation between the corresponding elements. Ehrig (2007) points out differentiations, claiming that mapping is a function between two ontologies represented by axioms that describe how to express concepts, relations or instances in terms of another ontology, focusing on the representation and establishment of the relations for certain tasks, whereas alignment only identifies the relation between ontologies.

Campos (2011) expresses a comprehensive definition when considering mapping as a formal expression that represents a function between two elements, equivalent or not, of distinct ontologies and allows the compatibility of such elements through actions for the execution of a determined task, being stored as a persistent model (which exists physically) and which can be separated or incorporated to one ontology. It is important to notice that both processes observed here have as a common characteristic to keep the original ontologies or vocabularies unaltered and in their places of origin but are able to generate links between these terms of these ontologies, expressing the type of relation that links the terms to be aligned.

Park (2006) and Boteram and Hubrich (2010) highlight the importance of the semantic issue in mapping initiatives. In this regard, we can notice here the relevance of Dahlberg’s (1983) proposal of the concept registry. DeRidder (2007), in turn, highlights the usefulness of a central vocabulary connecting vocabularies to be mapped. At this point, we can observe the relevance nowadays of the studies for the construction of intermediary languages and how the information science efforts could be explored to support concrete vocabulary compatibility initiatives. We consider here the relevance of citing a bridge to be considered between the semantic web trends, using the computer science view and the traditional practices of vocabulary control and knowledge organization and representation from information science. This bridge is SKOS, which is one of the most important semantic web specifications for application in archives, libraries, and documentation centers. It emerged as a W3C recommendation and offers a data model to represent the structure and content of conceptual schemas such as thesauri, classification systems and taxonomies. In its official documentation, SKOS points to the use of semi-formal knowledge organization systems to differentiate them from ontologies that support complex automatic reasoning processes, i.e., its orientation is towards indexing and retrieving information rather than more complex processing, which can be achieved, for example, with OWL (González 2014).

4.0 Conclusions

The increasing digital data volume produced in all areas of human acting, notably in the production of scientific data, requires urgent production of scientific data action by researchers and professionals, especially in the information science and computer science areas so that the data can be used effectively in an intelligent way to generate knowledge.

Therefore, today we have an imperative need to walk towards the creation of methods and techniques to produce semantic compatibility without necessarily being a strong human intervention since manual processes have already proven to be incapable of solving this problem. That is, we need to create requirements and models that allow the development of software applications able to align the heterogeneous vocabulary terms based not only on their codification formats but also on their semantic content.

We already have a number of techniques and technologies that allow the creation of interconnected data sources using resources such as universal resource identifiers, open linked data, etc., but on the other hand we have a huge and overwhelming set of data indexed by equally huge numbers of different vocabularies that represent local and peculiar idiosyncrasies that cannot be linked, whether due to technical, financial, administrative or even political difficulties, negatively affecting the retrievability of this information. In this sense, the growth and diversity of the web with its heterogeneous data sources, indexed by different vocabularies and distributed throughout different countries, companies and organizations, raises the need to discuss paths, requirements, technologies, guidelines and solutions that point to the possibility of automation considering the compatibility between different environments, not with the creation of a
new vocabulary to be used, because this is not always possible, but with the adoption of a metalanguage. This metalanguage must be formed as an intermediate language between the different source vocabularies enabling different actors to navigate this language and retrieve, contextually and semantically, the intended information no longer based on characters strings but on its meaning.

The techniques to implement the matching and compatibility of ontologies and vocabularies are complex processes that aim at reducing different representations, different perspectives and different modeling views. These techniques are a growing tendency since they are probably the best resources to encode the meaning of information, with the last decades being a period of extensive researches in this field. Nowadays these researches present an increase, and new publications where this problem is approached are continuously being published, reflecting the global interest in this matter (Otero-Cerdeira, Rodríguez-Martínez and Gómez-Rodríguez 2015).

Besides the proposals discussed here, we also verified the appearance, at the end of the last decade, of an investigation brought by the philosopher Pierre Lévy to elaborate a formal language called information economy metalanguage (IEML), which in a sense meets what we have been discussing so far. The IEML is proposal of a universal semantic addressing capable of indexing all the digital documents and has three basic procedures: 1) each distinct concept must have a single address; 2) the existence of a system of coordinated semantics must be open to any concept and relations between concepts (ontologies) regardless the cultural environment in which these concepts are created and transformed, without privileges and exclusions; and, 3) it must allow a group of mathematically defined operations (possible to be automated) on the semantic addresses, such as symmetries operations, logical inferences or comprehension, among others (Lévy 2007, 2009). The IEML semantic sphere presented by Lévy suggests a coordinated system where, first, the meanings are addressed and, then, represent movements and variations of the meanings through calculating functions. The IEML situates itself on the intersection between human languages and formalized languages and works as a categorizing system of all the culture, which results in development through digital means (Lévy 2014).

In this regard, we consider that in the investigation of such proposals and in the search for others it will be possible to identify methodological criteria that can allow an optimized appropriation considering the positive and negative points on the application of such methodologies in heterogeneous environment in which the scientific data present themselves.

References


Metadata*

Matthew S. Mayernik

National Center for Atmospheric Research (NCAR), University Corporation for Atmospheric Research (UCAR), P.O. Box 3000, Boulder, CO, USA, 80307-3000, <mayernik@ucar.edu>

Matthew S. Mayernik is a project scientist and research data services specialist in the National Center for Atmospheric Research (NCAR) Library, based in Boulder, CO, USA. His work is focused on research and service development related to scientific data curation. His research interests include metadata practices and standards, data citation and identity, and social and institutional aspects of research data. He received his master’s degree in library and information science and PhD from the UCLA Department of Information Studies.


Abstract: Metadata in various forms pervades our institutions, technologies, and daily lives. Metadata is a distinct focus of academic research and professional practice for many people within the library and information sciences (LIS). This article is an exploration of the concept of “metadata.” It presents a high-level introduction to the topic with analysis of key research problems and practical challenges. The paper discusses varying understandings of what “metadata” means, the origin and evolution of metadata as an important topic within information and data fields, and the central characteristics of that which gets called “metadata.” Metadata can be understood as both process and product and can result from both human effort and computational techniques. Given the central role metadata have in the establishment of knowledge, evidence, and truth, it is necessary for researchers and professionals within LIS to think critically about our metadata practices and systems.

Received: Revised: Accepted: 22 April 2020

Keywords: metadata, data, information, systems

* Derived from the article of similar title in the ISKO Encyclopedia of Knowledge Organization, Version 1.0, published 2020-03-16. Article category: Standards and formats for representing data. The author wishes to thank ISKO Encyclopedia of Knowledge Organization Editor-in-Chief Birger Hjørland for the invitation to write this article and for comments on previous versions as well as the anonymous peer reviewers for very insightful comments. This material is based upon work supported by the National Center for Atmospheric Research (NCAR), which is a major facility sponsored by the US National Science Foundation (NSF) under Cooperative Agreement No. 1852977. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author and do not necessarily reflect the views of NCAR or the NSF.

1.0 Introduction

“The future province of metadata is grand” (Greenberg and Garoufallou 2013, 2). For many people within the library and information sciences (LIS), metadata is a distinct focus of academic research and professional practice. LIS is unique in putting such a lens on metadata as a matter of disciplinary emphasis, but as indicated by the epigraph above, the scope of the people and institutions who are interested in, or work with, metadata is indeed grand.

Outside of LIS, metadata has traditionally been a prototypical infrastructural phenomenon: essential yet mundane and ubiquitous yet often invisible (Borgman 2003; Edwards 2010; Pomerantz 2015). In the past decade, however, metadata has emerged as a critical topic in many contexts. Metadata became a topic of political and legal intrigue with the publishing of stories about the US National Security Agency eavesdropping on digital communications (Schneier 2014; Mayernik and Acker 2018) and the use and manipulation of metadata gathered by social media platforms (Acker 2018). Online streaming services for music, movies, and other forms of personal entertainment rely on metadata of various kinds to provide recommendations, personalization, and categorizations to their users (Madrigal 2014;
of the characteristics discussed below manifest across a range of technical infrastructures and institutions.

2.0 “Metadata” within library and information science

Although the exact origins of the term “metadata” have been recounted in different ways over the past couple of decades (c.f. Greenberg 2005; Giles 2011; Gartner 2016), the generally accepted view seems to be that the term originated in the late 1960s in the context of computer system design to refer to the use of one data element to describe or represent some characteristic of another data element. A search of the Web of Science citation indexes in December 2019 shows that usage of the term “metadata” first appears in 1982 with rare and idiosyncratic usage through the 1980s. It started to become a term of niche usage in the early 1990s in discussions of information management systems, geographical information systems, and database design. Aside from sporadic early use, the term entered the discourse of the library and information sciences in the mid-1990s, particularly in relation to the development of digital library systems and the emergence of the internet and the World Wide Web as major social forces.

The term “metadata” became widely used during the mid-1990s to refer to approaches to information description, management, and discovery that differed from conventional cataloging approaches using library-focused structure and content standards like Machine Readable Cataloging (MARC) and the Anglo-American Cataloging Rules, Second Edition (AACR2). The development of the Dublin Core metadata element set in 1995 exemplifies this turn toward “metadata” within the LIS communities, both in concept and terminology. The Dublin Core, so named because it was formulated in a workshop in Dublin, Ohio, in March of 1995, was explicitly motivated by a desire to develop a common approach to describing electronic resources that would enable better discovery and collection of resources on the web (Weibel, 1995; Sugimoto, Baker, and Weibel 2002). Keeping track of web pages and other digital information resources with traditional library cataloging practices proved to be difficult, despite the best efforts of library professionals, because of the malleable nature of internet-based materials.

<table>
<thead>
<tr>
<th>Metadata subtag</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>A name given to the resource</td>
</tr>
<tr>
<td>Creator</td>
<td>A person primarily responsible for making the content of the resource</td>
</tr>
<tr>
<td>Subject</td>
<td>The topic of the content of the resource</td>
</tr>
<tr>
<td>Description</td>
<td>An account of the content of the resource</td>
</tr>
</tbody>
</table>
The Dublin Core element set included thirteen fields (later expanded to fifteen by including Description and Rights), as stated by Weibel (1995, n.p.) to be “the minimum number of metadata elements required to facilitate the discovery of document-like objects in a networked environment such as the Internet. The syntax was deliberately left unspecified as an implementation detail. The semantics of these elements was intended to be clear enough to be understood by a wide range of users.”

Weibel’s quote displays a couple of important points of debate that existed at the time (Lagoze, 1996) and continue to manifest in relation to metadata developments today.

- First, there is tradeoff in the extent of the metadata that is needed (e.g., minimal vs. comprehensive description) in relation the goals of the effort (document discovery in the case of the Dublin Core).
- Second, metadata initiatives often face challenges in defining an appropriate degree of standardization. In the case of the Dublin Core, it solved a certain kind of interoperability challenge by standardizing the names of the metadata elements but opened new interoperability challenges by not specifying the syntax of the information held by those elements.
- Third, in declaring that the Dublin Core was intended to be clear enough to be used by “a wide range of users,” the developers were explicitly going against prevailing approaches in which metadata standards and practices were targeted towards professional experts.

The Dublin Core thus exemplifies how the move toward “metadata” assumed and asserted that metadata descriptions for resources in the web environment would be created by a range of individuals, from expert to novice.

This last point particularly illustrates how early discussions explicitly centered on the ways that “metadata” existed as a counterpoint to conventional approaches to library cataloging (Greenberg 2005). Michael Gorman, editor of the Anglo-American Cataloging Rules (AACR) for many years, was a noted critic of the move toward metadata (Gorman 1999). In a later article simply titled “Metadata Dreaming,” Gorman (2006) stated that the approach to metadata development and implementation exemplified by the Dublin Core was based on a failed utopian dream of a “third way” of description (with the bibliographic description approach and the free-text “Google search” approach being the other two ways). In a subsequent memoir, Gorman has called metadata “an inferior, unstandardized species of cataloging done by amateurs” (2011, 191) that is targeted towards the “philosopher’s stone of bibliography—high-quality cataloging with no or little expense” (203). Gorman was perhaps one of the more visible and vocal critics but was by no means the only voice that argued that metadata projects that lose the more structured and complex approaches used in the library and archival community would struggle to be successful over the long term (Howarth 2005).

Such criticisms failed to slow the momentum of the “metadata” movement. The Dublin Core itself is the center of a dynamic metadata research and application community and is now a common reference point for many approaches to define “minimal metadata sets” for various purposes (Ara+kaki et al. 2018). It was also integrated into the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) as a way to facilitate harvesting of resources from diverse online sources (Van de Sompel et al 2004). The tensions noted above regarding metadata completeness, standardization, and consistency have not disappeared (c.f. Lagoze et al 2006; Urban 2014), but it is fair to say that the trends that motivated the development of the Dublin Core have held true. Namely, as most information and data systems have moved online and new types of internet-based information and communication technologies have emerged, the numbers of people, information types, and standards at play in the metadata space have increased correspondingly (Lagoze 2010).

Table 1. Dublin Core metadata standard.

<table>
<thead>
<tr>
<th>Metadata subtag</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Publisher</td>
<td>An organization or person responsible for making the resource available</td>
</tr>
<tr>
<td>Contributor</td>
<td>A person responsible for making contributions to the content of the resource</td>
</tr>
<tr>
<td>Date</td>
<td>The date that the resource was published or copyrighted</td>
</tr>
<tr>
<td>Type</td>
<td>The nature or genre of the content of the resource</td>
</tr>
<tr>
<td>Format</td>
<td>The physical or digital manifestation of the resource</td>
</tr>
<tr>
<td>Identifier</td>
<td>A string or number used to uniquely identify the object, e.g., the object identifier (OID)</td>
</tr>
<tr>
<td>Source</td>
<td>A reference to a resource from which the present resource is derived</td>
</tr>
<tr>
<td>Language</td>
<td>The language of the intellectual content of the resource</td>
</tr>
<tr>
<td>Relation</td>
<td>A reference to a related resource</td>
</tr>
<tr>
<td>Coverage</td>
<td>The extent or scope of the content of the resource</td>
</tr>
<tr>
<td>Rights</td>
<td>Information about rights held in and over the resource for rights management</td>
</tr>
</tbody>
</table>

The term “metadata” is now generally used in an expansive fashion to refer to descriptive and organizational schemes and practices broadly, regardless of whether they take place within information and data institutions or in other contexts. Like
its close relative “data,” the term “metadata” has been used to function as a plural or a collective singular noun (Rosenberg 2013). “Metadata” is commonly used now as a blanket term for a range of practices, many of which pre-existed the term “metadata” itself, including library cataloging, archival description, and scientific data documentation, along with more recent phenomena, such as automatically generated information associated with digital images or social media streams (c.f. Pomerantz 2015; Gartner 2016; Haynes 2017).

3.0 Metadata definitions, conceptions, and relations

As the term “metadata” has spread, it has been defined and redefined in numerous ways. Many scholars and professionals have moved past the most common definition of metadata, the literal “data about data,” to more nuanced and pragmatic discussions of requirements and functions.

3.1 Definitions

The following list provides a handful of definitions of metadata to illustrate how such definitions range from fairly specific to quite broad.

- Greenberg (2003, 1876)—“structured data about an object that supports functions associated with the designated object.”
- Greenberg (2005, 20)—“data attributes that describe, provide context, indicate the quality, or document other object (or data) characteristics.”
- Smiraglia (2005, 2)—“structured descriptors of information resources, designed to promote information retrieval.”
- Gilliland (2008, n.p.)—“the sum total of what one can say about any information object at any level of aggregation.”
- Pomerantz (2015, 26)—“Metadata is a statement about a potentially informative object.”

Within particular application areas or academic communities, more targeted definitions appear as, for example, in the following set of definitions of metadata by experts in geospatial and environmental science data.

- Michener, et al. (1997, 331)—“higher level information or instructions that describe the content, context, quality, structure, and accessibility of a specific data set.”
- Fegraus et al. (2005, 159)—“the information that describes ‘who, what, where, when, why, and how’ an ecological data-set was collected.”
- Danko (2012, 360)—“data that describes the information so that it will be useful and have value, be understandable, and enable collaboration.”
- Gordon and Habermann (2018, 38)—“well-defined content in structured representations that make it easier to share and discover.”

Jonathan Furner recently demonstrated how definitions of metadata also vary within standards established by ISO, the International Organization for Standardization (Furner 2020). Furner found that 96 ISO standards include definitions of “metadata,” ranging from “data about data” to much more detailed definitions. Furner concludes that while one interpretation of these findings is that the ISO standards represent a problematic inconsistency in what “metadata” refers to within the information and data worlds, another interpretation is that these varying definitions represent community-centric interpretations of the “metadata” concept as appropriate for their applications. Furner (2020, E41) states:

Perhaps it is actually more important for different domains, or even subdomains as represented by individual standards, to develop and to record their own particular definitions of terms, and thereby to make explicit the otherwise possibly overlooked differences in the ways in which the same terms are used in different contexts by different groups of specialists for different purposes.

One may argue that definitions in standards should take precedence over literature-based definitions. These two types of documents, however, tend to reach different audiences and, therefore, serve different purposes. Standards tend to be read and used by professionals while literature tends to be read by scholars and students. Therefore, neither source is more definitive. Rather, specific genres of documents (and specific individual documents) are more prominent within specific institutional situations. It is true that standards-based definitions are often developed by committees of individuals representing a variety of stakeholders. As Furner’s study shows, however, differences between standards-making bodies and committees themselves are a source of variation among definitions of “metadata.”

Definitions in both the research literature and standards, however, tend to focus on the use of metadata. This leads to the next topic, specifically, the ways that “metadata” have been categorized and conceptualized.

3.2 Categorizations and conceptions

If the notion of “metadata” has been defined in a variety of ways, it has been categorized and conceptualized in an even more diverse fashion. Categorizations of metadata reflect the different conceptions and motivations of the people who generate them and manifest in a variety of metadata ty-
Gilliland (2008), for example, in an overview article on metadata for library and information professionals, breaks the term “metadata” down into five types: administrative, descriptive, preservation, use, and technical. Other works, however, present different categorizations. A full explication and comparison of all of the categorizations that have been proposed is beyond the scope of this article, but looking across a selection of metadata-focused works (Greenberg 2001; Greenberg 2005; Lawrence et al 2009; Pomerantz 2015; Gartner 2016; Habermann 2018), the following categories appear at least once:

- access,
- administrative,
- archive,
- authentication,
- browse,
- character,
- descriptive,
- discovery,
- finding,
- identification,
- linking,
- preservation,
- provenance,
- relationships,
- rights,
- structural,
- technical,
- understanding,
- use.

Clearly some of these terms are related. But the variety of these categories indicates the generally broad understanding of what “metadata” might encompass.

Many of these categories represent particular tasks or actions that might be facilitated by metadata, such as authenticating, browsing, discovery/finding, preserving, or understanding information/data resources. This is perhaps the one commonality among these various definitions and categorizations of metadata: that metadata is created to be used, for some purpose(s), by people or computer applications. Karen Coyle (2010, 6) outlines how metadata is constructed, constructive, and actionable:

- Metadata is constructed: It is an artificial creation not found in nature.
- Metadata is constructive: It is created for a purpose, activity, or to solve a problem.
- Metadata is actionable: It is intended to be useful in some way.

Richard Gartner (2016, 4) provides a useful summation that encompasses this conceptualization of metadata as being designed and implemented for the purposes of particular uses:

The shape of metadata is designed by human beings for a particular purpose or to solve a particular problem, and the form it takes is indelibly stamped with its origins. There is nothing objective about metadata: it always makes a statement about the world, and this statement is subjective [in] what it includes, what it omits, where it draws its boundaries and in the terms it uses to describe it.

These characteristics of metadata hold across technologies, institutions, and decades. As one example, María Montenegro (2019) illustrates how the design of the Dublin Core metadata schema reflects the cultural assumptions of the people who were involved in its creation, particularly around notions of authorship and ownership of information resources. Information and knowledge that originate in other cultural contexts, such as within Indigenous communities, may not fit within the Dublin Core’s framework. As Montenegro (2019, 737) notes:

Two DC [Dublin Core] elements in particular perpetuate colonial practices of exclusion. Specifically, the rights and Creator Fields conflict directly with Indigenous epistemologies and protocols defining the access, circulation and use of TK [Traditional Knowledge] … Both fields—rights and creator—are formed upon and replicate legal frameworks that have embedded relations of exclusion. The definition provided by DC for the rights element presumes that IP [Intellectual Property] laws are universal, however, legal regimes of IP and copyright are culturally specific and the types of rights they specify, by definition, exclude all types of Indigenous TK.

In another example, Fidler and Acker (2016) depict some of the decisions that were at play in the design of the protocols for information exchange that underlie the internet. The designers of the internet protocols engaged in a range of debates about what metadata needed to be associated with each “packet” of information that was transmitted over the network. Discussions took place about the importance of socket numbers, network addresses of the computers at each end of the transmissions, as well identifiers for specific computer processes that were to be invoked by the transmissions. Other pieces of metadata were discussed but ultimately not included in the protocol’s requirements, including metadata related to the specific users who were doing the transmissions. These discussions were targeted toward
particular purposes, ranging from the technical functionalities that were desired to the need to potentially gather information for billing users for their usage of the network.

Such metadata can be found in the design of any networked information system; indeed, such systems cannot function without internal metadata that support the networks’ communications and functions (Mayernik and Acker 2018). This makes problematic the notion of metadata (or data) as being “exhaust” within technical systems. This “exhaust” metaphor has become increasingly common in discussions of metadata within digital systems (c.f. Mayer-Schönberger and Cukier 2013; Schneier 2015; Edwards 2017). Pomerantz (2015, 126) for example, states that “Up to this point, ‘metada’ has meant data that was created deliberately; data exhaust, on the contrary, is produced incidentally as a result of doing other things.” As we see from the Gartner quote above and the example from Fidler and Acker, any metadata created automatically within digital/networked information systems is a designed feature. There is nothing incidental about its creation. Using such metaphors as “exhaust,” “smog,” or “waste” when talking about metadata “implies that these traces are inevitable, a by-product of human and technical activities that cannot be avoided, and once produced are out of human control” (Mayernik and Acker 2018, 178). These metaphors serve to obscure understanding of metadata rather than illuminate it.

3.3 Relation to other concepts

This section provides brief overviews of how the concept of “metadata” relates to other important concepts within the library and information sciences.

3.3.1 Data

If we take the literal “data about data” definition of metadata, then it is straightforward to say that “metadata” is simply a sub-category of “data.” This is useful in that it allows us to characterize metadata as having certain properties that prior analyses have denoted for things classed as “data.” Two definitions of “data” are provided here as illustrations. Furner (2017, 66) defines data as “concrete instantiations of symbolic representations of descriptive propositions, informed by empirical observation, about the quantitative and qualitative properties of real-world phenomena.” Hjörland (2018) builds on Kaase (2001, 3251) to provide a more streamlined and generalized formulation: “Data are information on properties of units of analysis.” Both definitions note that “data” refers to entities that represent or contain information (“symbolic representations of descriptive propositions” in Furner’s terms) about other entities, whether “quantitative and qualitative properties of real-world phenomena” or “units of analysis.” Put more simply, Furner refers to data as “representational concreta,” that is, something concrete (i.e., materially manifesting via some real-world entity) that represents something else.

In this sense, it is straightforward to consider “metadata” to be a kind of “data.” The idea that “metadata” refers to “a statement about a potentially informative object” (Pomerantz 2015, 26), or other more specific definitions given above, fits well within the broad category of “data” as “representational concreta.”

Simply considering metadata to be a sub-class of data in this way is unsatisfactory, however, in that it does not provide any insight into why one might call a given entity “metadata” instead of “data.” Common distinctions within KO contexts, where metadata such as classifications or subject terms are greatly distilled representations or descriptions of informational resources, do not hold in some technical systems where the “metadata” stored by the system can be much bigger than the “data” (Klensin 1995; Brunton 2016).

Additionally, many pieces of information that are conventionally called “metadata” are in fact used by researchers and professionals as evidence to make particular claims. Think of the field of bibliometrics, for example, or the recent discussion of “bibliographic data science” by Lahti, et al (2019). Using something as evidence for specific claims is a key definition of “data” in the context of scholarly research, according to multiple recent scholars (Borgman 2015; Leonelli 2016). Mayernik (2019) argued that the distinction between data and metadata may be related to what is foregrounded and what is backgrounded in the context of a knowledge claim. In particular, metadata, “however instantiated in local situated activities of scientific research, are central to enabling something to serve an evidentiary role, that is, to serve as data. In particular, if data are entities used as evidence, then metadata are the processes and products that enable those entities to be accountable as evidence” (Mayernik 2019, 734-735, italics in original).

In sum, calling something “metadata” as opposed to “data” is a culturally contextual classification that rarely has a self-evident rationale (Boelstorff 2013). Data and metadata are often designated in contradistinction to each other, depending on the specific situations of origin and use (Borgman, Wallis, and Mayernik 2012; Mayernik and Acker 2018).

3.3.2 Document

The notion of a document is central to library and information science. Library and information work, including knowledge organization, centrally involves the creation, processing, and organization of documents. Library and information science scholars have thus developed sophisti-
cated understandings of what it means to call something a “document.” Michael Buckland (1997; 2014) outlines how particular entities can be “made as,” “made into,” and “considered as” documents. These three views, which are progressively more inclusive, reflect how: 1) particular things may be deliberately designed to serve documentary purposes (“made as documents”); 2) human artifacts may be used as documentary resources even if that was not their original purpose (“made into documents”); or, 3) naturally occurring objects such as rocks or animals may be used for documentary purposes (“considered as documents”). In these senses, almost any object could be used as a document depending on their evidentiary value in particular circumstances. Being a “document” is, therefore, a role that particular things play, rather than an inherent property of those things.

Furner (2016, 303) argues that all datasets are documents, stating that “the dataset is a species of document.” Thus, if metadata are a special kind of data, as noted above, then metadata likewise exist as documents not abstract concepts or information that exists without material form. As such, metadata can be analyzed via the same conceptual apparatus as documents. See Buckland (2018) for a recent overview on this topic.

If metadata is a sub-species of both data and document, it might be worth asking the question about the relationship among all three concepts. In other words, using Venn diagram terminology, one possible view is that the three terms are completely hierarchical, with “metadata” being completely encircled by “data,” which is in turn completely encircled by “document.” An alternate view is that “data” and “metadata” are partially overlapping circles within the larger “document” set. This latter view is a better fit with the discussion in the previous section of the culturally contextual ways in which data and metadata are distinguished. In other words, documents can be data, metadata, both, or neither depending on their usage as such in particular situations. Stated more concretely in relation to Buckland’s conception of “documents” as being roles rather than properties, being metadata is a role that some documents (or types of documents) have in particular circumstances (Renear and Wickett 2010).

3.3.3 Context

A couple of the definitions of metadata presented in section 3.1 referred to metadata as describing or providing context for informational/data resources. “Context” is itself a potentially slippery concept, generally referring to the setting or situation in which an action or event takes place and the factors that influence the action or event as it happens. Contexts can be important in how metadata is designed or in how it is interpreted (Wickett 2015). Dervin (1997, 14) notes that context is typically conceptualized, “usually implicitly, as a kind of container in which the phenomenon resides.” Talja, Keso, and Pietilainen (1999, 754) approach context from a metatheoretical viewpoint, saying “context is the site where a phenomenon is constituted as an object to us.” They describe context as the “crossroads between researcher and data.” Dourish (2004) describes how what is usually referred to as “context” can be better conceived as being rooted in “practices.” Shifting from “context” to “practices” allows us to focus on the “engaged action around artefacts and information that make those artefacts meaningful and relevant to people” (Dourish 2004, 26). Using this view, “context” exists both: 1) independently from the situated actions; and 2) co-produced by people via their situated actions. Metadata thus serve to create the context around information/data resources as much as they serve to describe that context.

As more metadata is produced automatically via computing systems, context is something that “must be reckoned in both architectural and institutional terms” (Agre 2001a, 194). In other words, “context,” in the context of computing systems, includes considerations of both the operations of the computing hardware and software—from bit-level to infrastructure-level—as well as considerations of institutional settings in which those computing systems were designed, created, and operated. Metadata associated with digital objects may be designed to reflect different parts of these details, depending on the application or situation.

4.0 Characteristics of metadata

As the diversity of the definitions, functions, and roles given above illustrates, metadata is not a definite and singular concept. Rather, it is a fluid, multiple, and fractional concept (Law 2004). Metadata is “fluid” in that file naming conventions, catalog records, data descriptions in repositories, user tags on YouTube, notes in personal Excel spreadsheets, email headers, and HTML tags can all be called “metadata.” Metadata, as a concept, is also characterized by “multiplicity” in that it is enacted differently in different social settings and situations, from Dublin Core records created by information professionals to descriptions in lab notebooks created by scientists to document their data. Despite this diversity, some characteristics and points of debate are common across metadata of different kinds. This section discusses some of the central characteristics of that which gets called “metadata.”

4.1 Structured vs. unstructured

A primary point of distinction in some discussions of metadata is between structured and unstructured information. A number of the definitions quoted in Section 3.1 explicitly
call out metadata as being “structured.” Many structures for metadata have been formalized into standards, ranging from general purpose metadata standards such as the Dublin Core to discipline-specific standards for particular kinds of resources, e.g., geospatial information (Danko 2012; Brodeur et al 2019). Standardized schemas and structures facilitate the interoperability of metadata between systems and applications (Zeng 2019). Metadata standards are commonly organized around a set of elements (such as “title,” “author,” “date”) that manifest as computer-readable documents in one of an alphabet-soup set of formats and markup languages, such as MARC, XML, JSON, and YAML.

Structured metadata can be differentiated from other forms of unstructured metadata, which might also be called “documentation” (Habermann 2018). Unstructured metadata could include any range of traces and practices that achieve some or most of the same goals as structured metadata, namely to create documentation, descriptions, and annotations for the purposes of managing, discovering, accessing, using, sharing, and preserving informational/data resources. As one example, in the context of data archives, it is common to include one or more narrative documents that describe various aspects of the data in more detail than is possible through standardized metadata structures.

It is important to keep in mind that structured and unstructured metadata can be hard to fully disentangle. Metadata standards commonly include a mix of controlled and uncontrolled elements. Controlled elements may require the information therein to conform to a specified syntax (e.g., “year-month-day” syntax in a date field) or to be chosen from a pre-determined set of values (i.e., controlled vocabularies). Uncontrolled fields, on the other hand, may allow any value to be present. Thus, even within highly structured metadata standards, there can be significant amounts of unstructured metadata. This characteristic can challenge attempts to aggregate or discover metadata, even if it is all structured according to a common standard (Arms et al 2002).

4.2. Metadata-as-product and metadata-as-process

The use of standards to create structured metadata results in what can be characterized as “metadata products.” Edwards et al (2011), in a discussion of metadata in the context of scientific research, describe how metadata products almost always involve corresponding “metadata processes,” namely, practices that help people overcome or bypass frictions that occur in the creation and use of metadata. Edwards et al (2011, 684) states:

Well-codified metadata products increase the precision with which a dataset can be fitted to purposes for which it was not originally intended, or can be reused by people who did not participate in creating it. At the same time, ephemeral, incomplete, ad hoc metadata processes act as lubricants in disjointed, imprecise scientific communication. This latter category of metadata frequently appears alone, in the case of datasets for which no metadata products exist, but it also frequently appears in the actual use of metadata products.

Some of the examples provided by Edwards et al (2011) and other related works (Mayernik, 2019) discuss how “metadata processes” effectively serve to facilitate data discovery, sharing, and use in situations where standardized “metadata products” have not or cannot be created due to time constraints or the lack of expertise available. As noted in the last sentence of this quote, however, metadata processes are also important in situations where standardized metadata products are being created. As one example, starting in 1988 and extending into the 2010s, the US Library of Congress published a set of “rule interpretations” for use by catalogers within libraries across the world who were creating catalog records via the Anglo-American Catalog Rules, 2nd Edition (AACR2). AACR2 provided hundreds of rules for cataloging library resources of all kinds. Applying these rules when cataloging particular items, however, involved interpretive decision-making regarding their fit to the details of the item in hand. The Library of Congress rule interpretations gave catalogers more detailed guidance on how to apply cataloging rules than was contained in the AACR2 cataloging code itself. These rule interpretations covered common cases, such as how to enter author names when there was more than one author of a resource and rare cases, such as how to designate authorship for a conference proceedings, where no individuals were named as authors or editors. As Barbara Tillett, former head of the Library of Congress’s cataloging division, noted (Tillett 2003, 113):

These rule interpretations lead to greater consistency in applying the rules, which is important for a very large institution and for its partners who help create compatible bibliographic and authority records. These guidelines are not appropriate for a cataloging code, but are needed for training and daily guidance to catalogers seeking to provide bibliographic description and access in a consistent way.

The Library of Congress Rule Interpretations (LCRI) were thus a kind of metadata process that facilitated the creation of more standardized metadata products. The implementation of any metadata standard is tied up in local interpretations and processes (Park and Mazaros 2009). This interpretive flexibility is a characteristic of every metadata standard or schema (Feinberg 2017). Looking closely at the pro-
duction of other metadata products would likely show similar couplings with attendant metadata processes.

4.3 Metadata and description

In a recent work, Michael Buckland (2017, 113) states that the “first and original use of metadata is to describe documents.” It is thus important to discuss briefly what “descriptions” are and what characteristics they pass on to metadata. The word “description,” like other similar words such as communication, illustration, and, yes, information, can be used to discuss both things and activities. When talking about metadata, descriptions are most commonly discussed as things, e.g., descriptions of library resources, archival materials, or data sets that are held in information systems and made accessible through catalogs. Decades of sociological research, however, has focused on description as an activity. This literature cannot be fully detailed here, but it provides important insight into how descriptions created and used as metadata should be understood.

Descriptions, whether verbal or written, are “only more or less reliable by virtue of their being treated that way for the practical purposes at hand” (Woolgar 1981, 509). In this sense, metadata encompasses negotiated shared meanings. Metadata is typically created with the expectation that readers or users of the descriptions will have knowledge of how to read and interpret them. As Heritage (1984, 150-151, italics in original) states, however, “no description is strictly compelled by the state of affairs it describes. Any description is thus inherently selective in relation to the state of affairs it depicts ... [C]hoices which underlie any description ... are all sources of clues concerning how the description is to be interpreted.” This characteristic, that metadata is inevitably selective, relates to the point in Section 3.2 above about metadata being created for specific purposes. Analysis of the metadata creation process should thus view metadata description, whether catalog records, classifications, labels, or technical traces, as a kind of action situated in social settings. In fact, as noted in the last sentence of the Heritage quote above, the metadata that does exist in some information system or social setting can itself be studied as a way to gather insight into the priorities, expectations, and accountabilities that exist in relation to those systems or settings (Mayernik 2019).

4.4 Search, discovery, and understanding

Beyond description, Buckland (2017, 118) notes that an additional use of metadata is to enable searching. Metadata can be used to provide structures that support consistent search and discovery of information across broad ranges of documents. Metadata can also potentially enable distinctions to be made among similar kinds of documents or resources. A search in a library catalog for “Hamlet” or a search in a scientific data catalog for “climate data” can result in hundreds or thousands of relevant results. Metadata that is useful for search and discovery may not be useful in distinguishing the differences among such large numbers of results. Users will likely need additional metadata that allows them to understand the resources, not just discover them (Habermann 2018). Providing metadata for understanding is certainly a role of the descriptions noted in the previous section. For example, annotated bibliographies of the various editions of Hamlet (Bevington 2019) and comparative guides for climate data (Schneider et al 2013) exist specifically to go beyond search and discovery to enable understanding. Metadata is not the only way to move beyond searching to understanding. Interface design and better search capabilities also have an impact (Marchionini 2006). But additional and novel metadata kinds and structures are central to this goal.

4.5 Relationships

One critical characteristic of metadata is that they are often the carrier of information about relationships within, among, and between informational/data resources. Many information and data systems manage and leverage relationships of a variety of kinds, including relationships among vocabulary terms and content structures (Bean and Green 2001) and relationships between documents and networks of documents (Mayernik 2018). Research in knowledge organization centers heavily on how to understand and represent relationships, both of a conceptual and a documentary nature (Green 2008; Szostak 2012) and has defined canonical types of relationships that obtain in the information arena, including hierarchical, associative, and equivalence relationships (Bean and Green 2001).

Yet, this aspect of metadata is often underappreciated. As Geoffrey Bowker (2016, n.p.) noted, “we don’t build our archives around relationships, we build them around things (if there is one fundamental flaw in our generic archival practices, it is this).” Gary Marchionini (2012), in an acceptance speech upon winning the “Award of Merit” from the Association for Information Science and Technology (ASIS&T) in 2011, suggested that “information science is in search of a theory of relationships,” (20) and stated that the community would benefit from paying attention “to the nature of relations in general rather than only identifying specific new relations” (21).

Rebecca Green (2008) discussed a variety of ways in which relationships manifest in knowledge organization systems. Relationships might be expressed via classification systems, vocabularies and thesauri, subject headings, or via specific relationship-focused elements in bibliographic records. Recent developments in the semantic web centrally in-
volve the precise specification of relationships between entities (Allison-Cassin 2012; Dunstire, Hillmann, and Phipps 2012). All of these manifest as metadata in some kind of document and/or information system. When relationship metadata exists in defined and structured form, they can be leveraged within information systems to enable information discovery and understanding, as well as to allow properties of one item to be transferred or inferred to another (Wickett 2018). When relationship metadata exists as unstructured information, e.g., as components of narrative metadata, they can enable keyword-based searching or be used by users to better understand the item(s) in hand.

5.0 Where does metadata come from?

As noted in the introduction, metadata comes from somewhere (Gitelman 2013). The social settings in which metadata are created have a large impact on what form metadata takes and on who or what creates metadata. Metadata can be created through both automated and manual processes. Both of these methods present challenges. This section outlines different people and technologies that have roles and responsibilities related to metadata creation.

5.1 Professional metadata creators

In libraries and archives, the creation of metadata is an institutionalized task. Catalogers, archivists, and professionals with titles like "metadata librarian" (Han and Hswe 2010) are assigned responsibility for creating metadata. Within these kinds of institutions, metadata work is also frequently conducted by paraprofessionals who have knowledge, experience, and training with the relevant systems and standards (Moulaison Sandy and Dykas 2016). Metadata creation is also a key responsibility for people working as data managers within data repositories (Palmer et al 2014; Rasmussen 2014).

Researchers and professionals from the library and information sciences (LIS) often approach their work through developing and applying defined sets of principles. Principles are discussed at professional meetings and in the literature, debated in standard-writing committees, and taught as part of professional education curricula. Principles offer directives for how information systems and the languages they use should be designed (Svenonius 2000). Principles depict how things should be or would be in optimal circumstances (Gnoli 2012).

The articulation of principles has been a central activity (and point of debate) within the library cataloging community for decades. Cataloging codes since the 1960s have been based in community-accepted principles, starting with the “Paris Principles,” which resulted from an international meeting held in Paris in 1961 (International Conference on Cataloguing Principles 1971). In the mid-1990s, when new cataloging code revisions were being debated, no fewer than three international conferences were held that focused either in whole or in part on the fundamental principles that should underlie the next code (Weihls 1998; Schottlaender 1998; Harkness Connell and Maxwell 2000). Individual contributions to these conferences debated the implementation of principles in past codes and presented new principles for a variety of specific issues, such as principles for cataloging relationships between resources and principles for cataloging serial materials. The cataloging code that resulted from these debates, titled Resource Description and Access, includes a statement of principles in the introductory chapter and notes at the beginning of each subsequent chapter how each section of rules relates to the stated principles (JSC 2014). The development of archival practices and institutions since the nineteenth century has been likewise driven by principle-based approaches (Gilliland 2014) as was the development of the Dublin Core Metadata Schema in the 1990s and early 2000s (Weibel 1995; Arakaki et al. 2018). Information and data professionals are far from having a monopoly on metadata creation, however, especially if the scope of what “metadata” entails is taken broadly.

5.2 Automatic metadata generation

As noted above, digital systems are inherently dependent on metadata that is automatically created for a variety of purposes (Mayernik and Acker 2018). The more structured the digital workflow, the easier it is to automate the creation of metadata, for example, to record provenance information about how information or data have been derived or changed over time. Beyond the use of automation to generate technical metadata, however, automation can also be applied to generate descriptive or topical metadata. Jane Greenberg (2004) described how automated metadata creation techniques typically follow one of two approaches, extraction or harvesting. In metadata extraction, “an algorithm automatically extracts metadata from a resource’s content” (Greenberg 2004, 62). Common applications of the extraction approach include automatic abstract generation for publications and summary displays of web pages given by web search systems. Metadata harvesting, on the other hand, involves compiling metadata automatically from distributed resources, such as collecting standardized metadata from metadata feeds or web site HTML. As Greenberg (2004, 63) notes, “the harvesting process relies on the metadata produced by humans or by full or semi-automatic processes supported by software.”

Automated metadata extraction and harvesting methods are most robust for text-based documents, for time- and geotagging of digital photos, and the like. But new technologies and techniques for extracting information from audio
recording, video, and images also have promise for the purposes of metadata creation (Riley 2017). Facial recognition software could be used, for example, to create metadata about the people that are shown in digital video or image collections maintained by libraries and archives. Given the explosion of digital media and the growth of digital archives, these kinds of techniques may be the only tractable way for such metadata to be produced (see for example Mühling et al 2019). The use of facial recognition and other similar algorithmic metadata extraction techniques must be coupled, however, with strong awareness of the notable ethical implications that arise when creating information about people without their awareness (Agre 2001b; Seeman 2012; Crawford 2019; Padilla 2019).

5.3 Metadata creation in everyday life

Outside of information/data institutions and structured technological workflows, metadata creation can take on various forms with many opportunities and challenges. Many of the computational techniques noted in the previous section are either in nascent form or are not effective when applied to unstructured or very diverse informational resources. They may also require specific technical skills to implement. In day-to-day life, people may create metadata for information and data in the context of work or home settings. People may create metadata via folder structures and file names for personal images, or create one-off notes documents for particular tasks or resources. All of these are acts of metadata creation in the general sense.

Internet-based tools for sharing photos, videos, and other kinds of information commonly enable users to add metadata as tags to text or objects. Tags are also common within social media systems, where users of Twitter, Facebook, or Instagram include hashtags to connect their posts with other discussions within the platforms, such as #WomensHistoryMonth, #EmployeeAppreciationDay, or #data. Library and information science researchers have studied how the aggregation of such tags can create folksonomies that reflect the vocabularies and language usage of everyday people in contrast to the structured and pre-determined taxonomies created and used by information professionals. Folksonomies have been studied and implemented as ways to bridge between expert and non-expert vocabularies (Cairns 2013) and potentially to feed into the creation of formal taxonomies or ontologies (Gil et al 2017). Such “crowdsourcing” of metadata has benefits and drawbacks. Enabling users to add metadata via their own terminologies tend to better support information browsing than searching (Sinclair and Cardew-Hall 2007) but can be very effective at accommodating and celebrating multiple voices and perspectives on the resources being described (Srinivasan et al 2009).

Metadata created for everyday tasks should be expected to have different characteristics than metadata created by professionals for institutional purposes. Such metadata tends to be idiosyncratic, varying in content and structure from individual to individual, and from situation to situation for the same individual. The aforementioned discussion of principles for metadata do not apply. As Chamberlain and Crabtree (2016, 569) note in a study of how metadata is created and used in the context of personal music collections:

Relevance is a key factor to understanding the nature of metadata, what is relevant in one context may rapidly change as different artifacts, reasons and results are employed in different emerging contexts – metadata is not always a static “entity,” in many respects it consists of different physical modalities, relates to people (trusted) and has different perceived and actual temporal qualities. Our fieldwork shows that the emergence and use of metadata is both part of, and yet can be separated from the workflow.

Creating metadata in everyday situations is often a task that has implicit or explicit moral implications (Vertesi et al 2016). People feel morally responsible for keeping track of important documents and information, such as family pictures, yearly tax documents, or vaccination records for children. These types of documents may exist in a variety of digital and analog formats and in multiple technical environments (e.g., email, personal computers, mobile phones). Creating metadata and organization systems for these kinds of things can thus be a source of emotional and interpersonal stress. People thus make context-specific choices about what metadata to create and in what forms and systems.

5.4 Metadata collaborations

Metadata creation often involves collaborations between people with varied expertise and knowledge. People with special domain or disciplinary knowledge may seek out metadata experts for help in the creation of specialized collections of resources on a particular topic. Or libraries and archives may bring in experts on a particular topic to provide consulting or specialized knowledge for particular collections. Additionally, as more libraries and archives are collecting born-digital resources, such as digital datasets, software packages, and other kinds of materials, they often have multi-step workflows where the contributor of the resource is required to create metadata for their asset(s), with professional librarians or metadata experts providing review and quality control for the resulting metadata.
Such collaborations can be challenging. People with different perspectives will bring different expectations about technologies, workflows, and outcomes (Khoo and Hall 2013). The time, energy, and attention involved in creating, collecting, assembling, checking, and/or understanding metadata can be significant, particularly for people without experience in creating structured or non-structured metadata. For example, scholarly research data repositories commonly experience difficulty in getting data creators to create metadata, and the metadata that is created can be of minimal quality (Jones et al 2006; Bhandary et al 2018). In some cases, researchers will refrain from sharing data, because it takes too much effort to produce the data and associated documentation necessary for its use (Tenopir et al 2011).

The benefits of such collaborations, on the other hand, center on being able to take advantage of the different sets of expertise that different people may bring. For example, disciplinary topic experts have firsthand knowledge about how resources related to their areas of expertise are created and used, and thus can provide useful insight into what metadata should be created, and how the metadata may be optimally structured to support use and re-use (White 2010). In the best circumstances, metadata experts and non-experts work together (Gazan 2003). Information and data professionals may serve as intermediaries (Mayernik 2016) to support the optimal usage of the applicable metadata standards, vocabularies, while topical experts provide relevant metadata content and guidance on usage.

6.0 Metadata futures—conclusions and research questions

This paper begins with a quote stating that the future of metadata is grand. The subsequent sections illustrated how metadata in various forms pervades our institutions, technologies, and daily lives. The ongoing digitization of our societies is, if anything, accelerating this trend. As noted by Richard Gartner (2016, 96):

The growth of the digital seems to need more metadata not less. Google and its peers make it possible to discover new material in ways which we could not have conceived of before but they need to be complemented by human thought and the metadata by which it is focussed.

Going forward, researchers and professionals will continue to grapple with long-lasting challenges related to metadata creation and use, including questions about how to negotiate cost/benefit trade-offs between structured and unstructured metadata and between human and machine generated metadata. But it is also clear that each new generation of information and data technology produce and require different kinds of metadata than systems that existed before. Social media technologies demonstrate a trend toward what Ronald E. Day (2019) has called “a posteriori” metadata generation. Day uses this term to contrast with “a priori” metadata generation as in library cataloging and classification, semantic web ontologies, and scientific data catalogs, where the generation of metadata tends to come at the beginning of the life cycle of information/data use. In systems based on “a posteriori” metadata, the metadata that get generated, stored, and used within the systems are less focused on the properties of the entities within those systems, instead focusing more on what those entities do (or what is done to them by other entities). Twitter, for example, may collect metadata about particular Twitter users, but they monetize metadata that reflects what those users do. Likewise, Twitter generates metadata about each post, but they monetize metadata that reflects how those posts travel (how many likes, retweets, and replies are generated), and the social networks connected to those interactions.

Information and data researchers and professionals who have expertise in the “a priori” metadata approach have much to consider in this trend toward “a posteriori” metadata. As Day (2019, 138, italics in original) suggests, “A priori categories, such as those that result from classification structures, can be heuristics for investigating entities, but they are only that.” What does knowledge organization consist of if the value and meaning of particular information and data resources is “based on statistical calculations of the use and the relations of data” (141) instead of a priori decisions about subjects, classes, and categorizations based on inherent properties of those resources? Open research questions on this point relate to the relative value of both approaches to metadata creation in relation to the goals of the metadata being produced. In section 3.2, I listed nineteen different types of metadata that have been identified in prior literature (administrative, descriptive, discovery, preservation, provenance, technical, etc.). Some of these metadata types are obviously conducive to the “a posteriori” approach, such as provenance metadata, but with other types, such as descriptive and discovery metadata, the relative value of the “a priori” vs “a posteriori” approach to metadata creation is still an open question.

Day’s reflections likewise provide important considerations for the future of metadata and knowledge organization in relation to evidence and evidentiaryness in the digital age. In a time of “fake news” and “deepfakes,” that which is considered knowledge is tied to the kinds of evidence that exist to buttress knowledge claims and to the ways that that evidence is marshalled. The use of documents as evidence has always been tied up with assessments of the authenticity of the documents involved. Authenticity assessments in the digital age rely on metadata to provide (and create) context and accountability for evidence. When the metadata nece-


Knowledge Organization

Seventh Annual “Best Paper in KO Award” for Volume 46 (2019)*

DOI: 10.5771/0943-7444-2020-8-714

Awarded jointly to:

Vanda Broughton, for “The Respective Roles of Intellectual Creativity and Automation in Representing Diversity: Human and Machine Generated Bias.”

Knowledge Organization 46, no. 8: 596-606.

and

Richard P. Smiraglia, for “Work.”

Knowledge Organization 46, no. 4: 308-19.

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.

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: A work is a deliberately created informing entity intended for communication. A work consists of abstract intellectual content that is distinct from any object that is its carrier. In library and information science, the importance of the work lies squarely with the problem of information retrieval. Works are mentifacts—intellectual (or mental) constructs that serve as artifacts of the cultures in which they arise. The meaning of a work is abstract at every level, from its creator’s conception of it, to its reception and inherence by its consumers. Works are a kind of informing object and are subject to the phenomenon of instantiation, or realization over time. Research has indicated a base typology of instantiation. The problem for information retrieval is to simultaneously collocate and disambiguate large sets of instantiations. Cataloging and bibliographic tradition stipulate an alphabetico-classed arrangement of works based on an authorship principle. FRBR provided an entity-relationship schema for enhanced control of works in future catalogs, which has been incorporated into RDA. FRBRoo provides an empirically more precise model of work entities as informing objects and a schema for their representation in knowledge organization systems.

Richard P. Smiraglia holds a PhD in information from the University of Chicago. He is Senior Fellow and Executive Director of the Institute for Knowledge Organization and Structure, Inc. and is Editor-in-Chief of this journal. He also is Professor Emeritus of the iSchool at the University of Wisconsin-Milwaukee. He was 2017-2018 KNAW Visiting Professor at DANS (Data Archiving and Networked Services division of the Royal Netherlands Academy of the Arts and Sciences), The Hague, The Netherlands, where he remains visiting fellow and was the 2018 recipient of the 2018 Frederick G. Kilgour Award for Research in Library and Information Technology.

* Awards Committee for Volume 45 (2018): Ann Graf, chair; Jane Greenberg, Joseph Tennis, Daniel Martínez-Ávila and Yejun Wu


