Knowledge Maps of the UDC: Uses and Use Cases†

Andrea Scharnhorst*, Richard P. Smiraglia**,
Christophe Guéret*** and Alkim Almila Akdag Salah*

*Data Archiving and Networked Services DANS, Anna van Saksenlaan 51,
2593 HW Den Haag, Netherlands, <andrea.scharnhorst@dans.knaw.nl>,
<http://www.narcis.nl/person/RecordID/PRS1285871/Language/en>,
<christophe.gueret@gmail.com>, <alelma@gmail.com>
** University of Wisconsin-Milwaukee, School of Information Studies,
2025 E Newport, Milwaukee, Wisconsin 53211, USA <smiragli@uwm.edu>
*** Broadcasting House, Llantrisant Road, Cardiff CF5 2YQ, Wales, UK,
<christophe.gueret@bbc.co.uk>

Andrea Scharnhorst is the coordinator of the Research & Innovation group at Data Archiving and Networked Services (DANS), Royal Netherlands Academy of the Arts and Sciences (KNAW), Amsterdam. With a background in physics and in philosophy of science, she has published models of innovation and science dynamics and edited a 2012 book, *Models of Science Dynamics,* with Katy Börner and Peter van den Besselaar. She has empirically studied the Matthew effect of science, analyzed the structure and emergence of knowledge orders, and written about the need for visual interfaces.

Richard P. Smiraglia is a professor and member of the Knowledge Organization Research Group in the iSchool at the University of Wisconsin-Milwaukee. He has explored domain analysis for evolution of knowledge organization, epistemological analysis of the role of authorship in bibliographic tradition, the evolution of knowledge and its representation in knowledge organization systems, and the phenomenon of instantiation among information objects. He is Editor-in-Chief of this journal.

Almila Akdag Salah studied at the Art History Department of the University of California, Los Angeles (UCLA), focusing on technoscience art and its place in the art historical canon. She was one of the first Digital Humanities Fellows of UCLA. Upon completing her PhD, she became a Postdoctoral researcher with the Virtual Knowledge Studio (KNAW)—Knowledge Space Lab project, which contributed to the new research area of maps of science. Almila joined the Digital Methods Initiative (University of Amsterdam, New Media Studies) with a Veni grant on the online art platform deviantArt. She is the Netherlands Management Committee member for the FP7 COST project KNOWeSCAPE on knowledge spaces. Currently she works as an associate professor at the Istanbul Sehir University's College of Communications.

Christophe Guéret is a researcher specialising in knowledge representation and reasoning and computational intelligence. He is currently working on knowledge integration systems at the BBC in Wales. He previously worked as a research fellow at Data Archiving and Networked Services (DANS) and the eHumanities group of the KNAW in the Netherlands. His research activities are focused the design of decentralised interconnected knowledge systems and the analysis of their content and social implications.

Scharnhorst, Andrea, Richard P. Smiraglia, Christophe Guéret and Alkim Almila Akdag Salah. 2016. "Knowledge Maps of the UDC: Uses and Use Cases." *Knowledge Organization* 43(8): 641-654. 27 references.

Abstract: Insight into the depth and breadth of knowledge for use in and across disciplines is of vital importance. Our knowledge maps are visualizations based on empirical evidence about both collection characteristics and knowledge clusters such as disciplines. We report in this paper on collaborative efforts over several years, combining the resources of the Knowledge Space Lab and the Research and Innovation Group at DANS. In particular, we were interested in the narrative of how knowledge and knowledge systems change over time.

Knowledge organization systems are evolving complex systems. Their analysis, both concerning inner structure, evolution over time, and their implementation in information spaces is important to better understand how knowledge is produced and can be navigated through.

We applied a mixed research method strategy to the analysis of the Universal Decimal Classification (UDC), combining web-based data collection with data and visual analyses. The growth of the UDC over the twentieth century parallels the evolution of knowledge in the academic canon. Rather than reconstruct main classes with potentially catastrophic revisions, the editors of the UDC preferred complex and ever more granular evolution of special auxiliaries. In evaluating the population of the UDC, we have seen even more evidence of the cultural evolution of knowledge across time. While this approach to research is important for knowledge organization as such, it also bears potential for information providers to use visualizations to showcase their collections.

† Originally created for: International UDC Seminar "Classification & Authority Control: Expanding Resource Discovery," in Lisbon, Portugal, 29-30 October 2015. The authors wish to thank Ed O'Neill of the OCLC Office of Research, Johan Rademakers and Bart Peeters of KU Leuven, and Maria Inês Cordeiro of the National Library of Portugal for UDC data analyzed for this research. Part of this work has been supported by the COST Action TD1210 Knowescape. We are indebted to Aïda Slavic for her professional and practical support throughout this research process.

Received: 2 April 2016; Revised: 23 August 2016; Accepted: 23 August 2016

Keywords: UDC, Wikipedia, classification, knowledge organization systems, evolution of ontologies

1.0 Insight into knowledge for use

Visuals are all around us and visual analytics is now embraced not only for business purposes but also as a research method applied across all sciences, including the humanities. Concerning the latter, the work of Lev Manovich and others is an example of how big data analytics, visual analytics, and art history meet in what has been termed "cultural analytics" (Salah, Manovich, and Crow 2013). But, when it comes to libraries and archives, those guardians for cultural heritage and public access to knowledge that long since have joined the world of automatically processed information, applications of visual analytics are rather sparse. Having said this, we also should say that the world of libraries has not been untouched by the metric wave. In the literature, we find indicator sets about the performance of libraries (Heaney 2009). Among them are some—like the size of the collection—which are also relevant in the light of our own analysis. However, our purpose is not primarily to analyze the institutional functions of the library. We are interested in providing insight into the depth and breadth of knowledge for use in and across disciplines. In other words, we want to know what kind of knowledge we find in a library or archive, and we use knowledge organization systems developed to create structure in and access to the content of collections to gain an overview about knowledge in them.

Our method is to produce visualizations from baseline statistics. These knowledge maps are based on empirical evidence about collection characteristics on the library and archives side, and knowledge clusters such as disciplines, for example, on the other.

When looking at knowledge organization systems (KOSs) as applied in and populated by collections, we cannot avoid also looking at the KO systems themselves. Contrary to naïve beliefs, classifications are not fixed, they evolve with the needs for which they have been developed and with the changing insights in the content for which they have been developed. For example, consider how Linneaus' system for classifying the biological species has developed since its inception (Ereshefsky 2001). The same sort of evolution takes place when it comes to ordering knowledge, as the history of scientific classification shows (Kedrov 1975-76). There, it holds that changes in classification are always as much as triggered by changes in the objects to be classified as in the changes of points of view of those who classify them. According to Bowker and Star (1999), the authoritative voice of classifications reflects the *Zeitgeist*. Classifications are a kind of social product, but because of their authority, which is amplified in impact when used for machine-based large-scale information processing operations, it is important to analyse changes. The analysis of changes should be presented both in detail and at a meta-level (e.g., Börner 2010 and 2015). Here, our work meets the few explorations into the evolution of KOSs from classification research (Tennis 2012; Tennis, Thornton and Filer 2012) and the relatively uncoupled parallel investigations into the evolution of ontologies (Noy and Klein 2004; Meroño-Peñuela 2016).

Thus, there is need for interdisciplinary work in this area. We report in this paper on collaborative efforts over several years, supported by different project grants. Our own team consists of computer scientists, physicists, information scientists, and humanists. Our research into the population of the Universal Decimal Classification (UDC) has generated immense quantities of specific data about particular classification attributions to and among particular and specific documents. Our goal is to learn how to create knowledge maps of these data. Knowledge maps, like geographic maps, serve different functions. Knowledge maps can serve as awareness catalysts, reference systems, data curation vehicles, and heuristic devices for research (Scharnhorst 2015).

In particular, we are interested in the narrative of how knowledge and knowledge systems change over time. We all are accustomed to the notion that things change, and we work hard to keep up with changes. This takes place in a forward direction. In the science of knowledge organization, it is important also to analyze the accretion of change over time in reverse. In other words, it is important not only to be able to make a map of knowledge today, but also to make maps of knowledge over time leading up to today. This is particularly challenging since both the coordination system that hosts the knowledge system and the classification and production of knowledge change at the same time. One of our goals is to make it possible to transit both forward and back through knowledge evolution. In this paper, we will describe one approach to such visualization based on the analysis of the evolution of the UDC.

1.1 The notion of a reference classification

Our story began with the Knowledge Space Lab (KSL), a research group within the Virtual Knowledge Studio of the Royal Netherlands Academy of the Arts and Sciences. KSL was a project begun in 2009 with the goal of creating a knowledge map of the evolution of science by tracking the evolution of knowledge in the then emergent Wikipedia. KSL downloaded the latest dump of the English Wikipedia that was available at that time (https://archive.org/details/enwiki-20080103). Wikipedia continues to publish data dumps, which include whole histories of every Wikipedia page. Not surprisingly, those dumps are growing in size. In the time of the KSL project, an alliance with powerful computer centers was needed to store and process the Wikipedia data. This is why we applied for a grant from BigGrid.nl that gave access to high computing. Thus the Wikipedia project was one of the few humanities projects that made use of grid computing. KSL extracted all changes of links in Wikipedia pages. The team was interested in the growth and change of the topical classification, and to this purpose extracted all changes of links between category pages and article pages. Similar to the collective editing of any page in Wikipedia (be it a category page, an article page, or another page type), the relationships between categories are debated. There exists one page in Wikipedia (2015) that demonstrates this (https://en.wikipedia.org/wiki/Category:Main_topic_classifications). Here, at the time of this writing, one finds fourteen subcategories; in 2008, one would have seen forty-three subcategories. However, while the page as such has a history, at any given moment those subcategories listed are dynamically created from the present Wikipedia. For example, one might visit the archived 2008 page to see the categories present at that time (https://en.wikipedia.org/w/index.php?title=Category:Main_topic_classifications&diff=240586527&oldid=23865691). This blind spot in Wikipedia concerning its own memory triggered the reconstruction by our team of all link relationships over time. The KSL team at the end provided monthly snapshots for both the network of page links and the network of category links, reconstructing the categorical network in Wikipedia (Suchecki et al. 2012). We explicitly encourage re-use of the data, and would like to remark that the monthly snapshots of the network have never been visualized nor fully analyzed.

The team analyzed the main topical classification where present over time, and how re-ordering of the category system is reflected in changes of the topology of the whole network of links between category and article pages. But, the team also wanted a control for their experiment, and it was decided that a stable bibliographic classification could provide that control. In other words, a bibliographic classification based on literary warrant—i.e., based on concepts found in the published academic canon—could be visualized alongside the Wikipedia to show the parallels and divergences in the evolution of knowledge. The Wikipedia was known at that time for rapid growth, if not so much for accuracy. On the other hand, bibliographic classifications are known for just the opposite—measured change over time and only once change has taken hold in the published literature of the academic canon.

While the Wikipedia data were churning on the grid, the KSL team set its sights on the UDC. Quickly, we learned there was no one UDC. Unlike the *Dewey Decimal Classification*, with which the UDC shares common origins, there was not a systematic set of editions published over time containing the whole classification, each edition enshrining change at a moment in time. Instead, the UDC has always been maintained as a virtual classification, with its entirety available only to its editorial board (McIlwaine 2007, 1-4). Individual chunks of it have been published in various languages at various times, but always with strict limitations on the depth of classes and the extent of granularity of subdivisions. In a sense, our team was given a bold opportunity to discover the entirety of the UDC as best we could and to then map its evolution over time. Our work in this vein was reported in several papers (Salah et al. 2012, Smiraglia et al. 2013, Scharnhorst and Smiraglia 2012a) and the eventual evolution was mapped alongside the evolution (then) of the Wikipedia in a now famous knowledge map that can be found online (http://scimaps.org/mapdetail/design_vs_emergence__127) and as part of the *Atlas of Knowledge* (Börner 2015).

To understand how this comparison works one has to be aware that the Wikipedia category system is a fully connected graph with cycles, and not a tree from a point as is the UDC as we know it in the form of its Master Reference File and the classes that can be represented. We emphasize this, because our 2013 paper (Smiraglia 2013) reported the task of reconstructing a network from the UDC, as we will discuss below. But for the comparison presented in the map we applied "brute force" and turned the Wikipedia network into a tree, just taking the "Main topic_classifications" page as the root and ignoring all back references from low level nodes to high level nodes. In Figure 1, we present the Wikipedia network on the left and the UDC network on the right. The comparison is possible because we use color-coding for matching categories. We started from the UDC, and then allocated the 43 high-level topical categories of Wikipedia in the UDC. This allocation also was a complex process, because term comparison gave only one indication. Moreover, terms can also appear at different levels, and eventually, terms can have a different meaning. In some cases, we manually inspected the related Wikipedia article pages to decide on the proper matching.

In the Wikipedia, you see the dynamics of the evolution of early 21st century thought. You see the interest in all kinds of art phenomena—pop artists, radio stations, films. In the case of the UDC, you see the concatenation of more than a century of academic canonization. They are not comparable. Rather, they are complementary. The stable reference system represented by the UDC is a record of the canonization of the evolution of knowledge over a century, and its application in libraries. The evolving system represented by the Wikipedia is a visualization of the dynamism of emergent thought and culture, replicative science, paradigm shifts, and all that goes into (eventually) the stability of the canonized reference system. In simple terms, UDC reflects the growth of academic knowledge (green represents sciences, and purple stands for arts and entertainment), whereas Wikipedia reflects contemporary cultural interests. This simple visualization is a true map at one point in time of (as it says) the emergence of knowledge orders. One has to admit that in the case of the UDC, we looked "only" at the category system, while in case of the Wikipedia, we see categories and how they are populated. Also, the Wikipedia in 2008 seems more balanced in some ways and so has more resemblance to Otlet's original classification. These small imperfections of the "design versus emergence" map motivated us to have a closer and deeper look into the UDC.

2.0 How the UDC has evolved over time

A primary objective was to create a narrative about the evolution of the UDC from 1905 to the present. Once

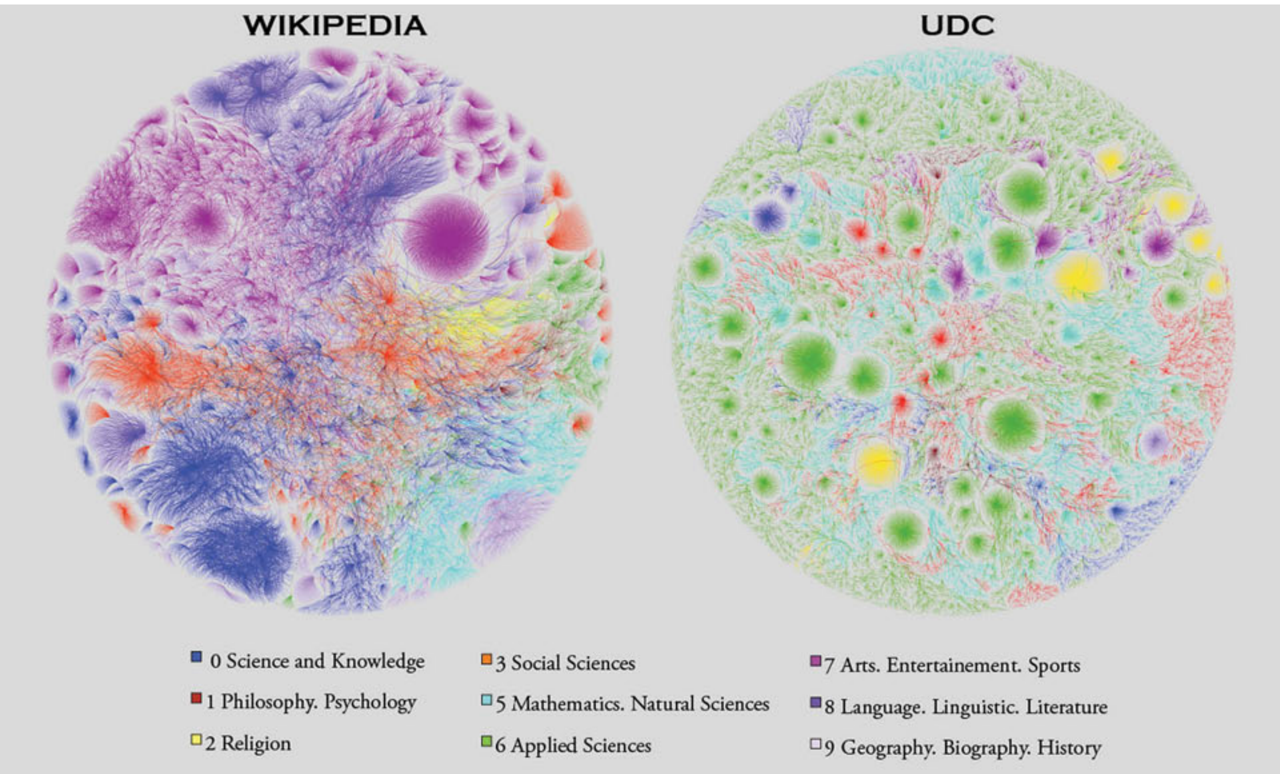

Figure 1. Visualizing Wikipedia and UDC networks.

the entire UDC had been digitized and mapped over time, and after the team gained access to the current Master Reference File, which is kept online, we were able to create detailed visualizations of the growth of classes and auxiliaries over time. Most of the details, including methodological steps, are reported in Salah et al. (2012), but a summary here will point to the efficacy of the technique. Our usual starting point is a visualization of the ten main UDC classes comparing first 1905 and 2005 and 2008. The changes between 2005 and 2008 are only incremental (Figure 2).

In Figure 2, the doughnut shows the main classes of the UDC in 1905 in the inner ring and 2005 in the middle ring and 2008 in the outer ring. The main observations are the decreased size of class "0 Generalities" and the increase in classes "5 Mathematics, Natural Sciences" and "6 Applied Sciences, Medicine, Technology." In 1905, the "0" class would have held mostly multi-volume reference sources; by 2009, the class has become "Science and Knowledge. Computer Science. Information." The immense granularization and growth of sciences during the twentieth century is reflected in both shifts. We believe also that this is a visualization of literary warrant—that is, UDC is based primarily on the growth of canonical literature in academic libraries and it is through that lens that we are able to view the growth of productivity in the sciences in terms of more and more granular literature.

Another visualization of increased granularity comes from analysis of the growth of auxiliaries over time. In the UDC, auxiliary schemes are used to express complexity through a process of synthesis. That is, a symbol from an auxiliary table is appended to a symbol from a main class to express a complex relationship. Common auxiliaries provide a form of facets to express form, time, place, language, ethnicity, etc. Special auxiliaries function in the same way but are limited to specific main classes. For example, main class "2 Religion" has changed little in size since the earliest iteration of the UDC, but the entirety of the coverage of religion was reworked predominantly as special auxiliaries to be added to a few main classes. Figures 3 and 4 show how auxiliaries have changed over time. For example, Figure 3 shows the changes in class "2" since 1998; in fact, 90.07% of class "2" is comprised of special auxiliaries post-1998. There clearly also is major and continuous evolution of special auxiliaries for class "6 Applied Sciences, Medicine, Technology."

Figure 4 shows the growth in granularity in auxiliary "e" (place names), with some sustained growth also in auxiliaries "c" (persons) and "k" (materials). Figure 4b shows how the common auxiliaries are distributed among the main classes in the Master Reference File; again we see that the sciences predominate, which is a reflection of

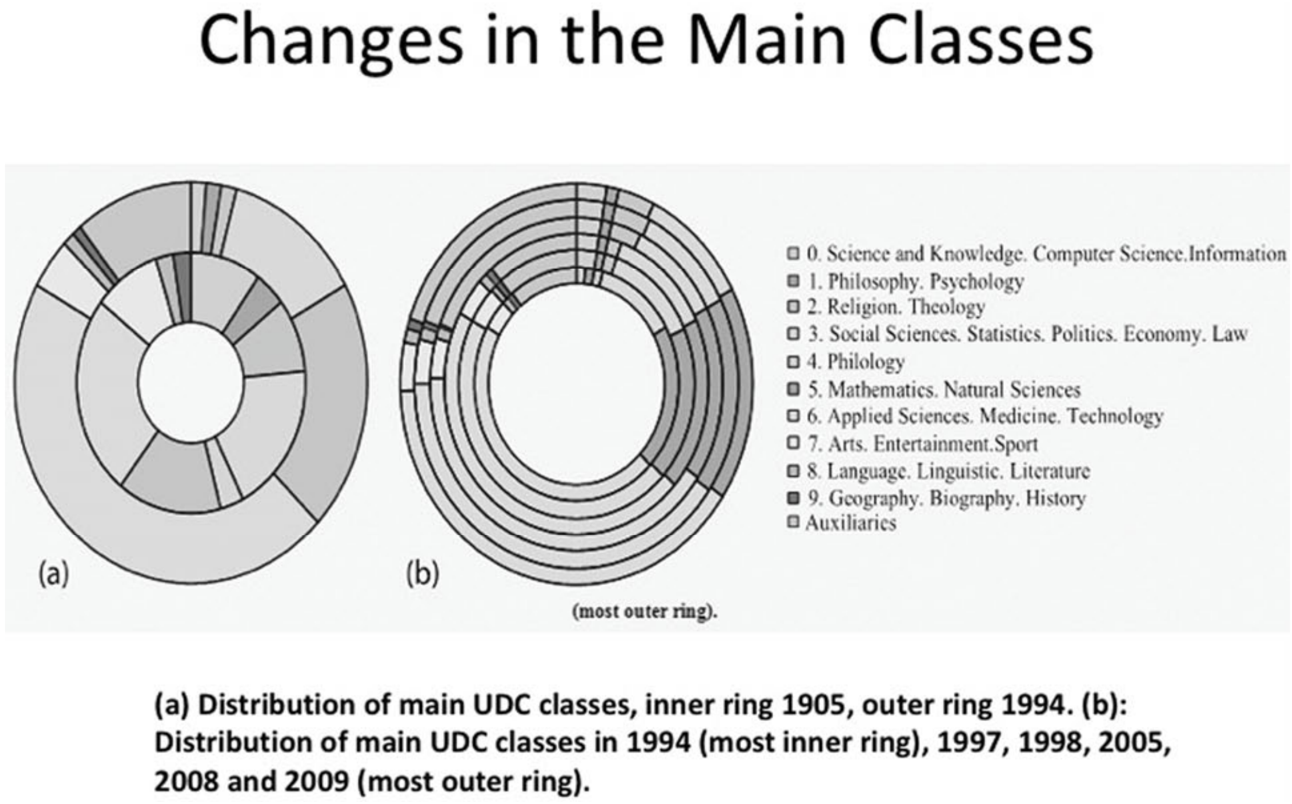

Figure 2. Changes in UDC main classes (Salah et al. 2012, 53).

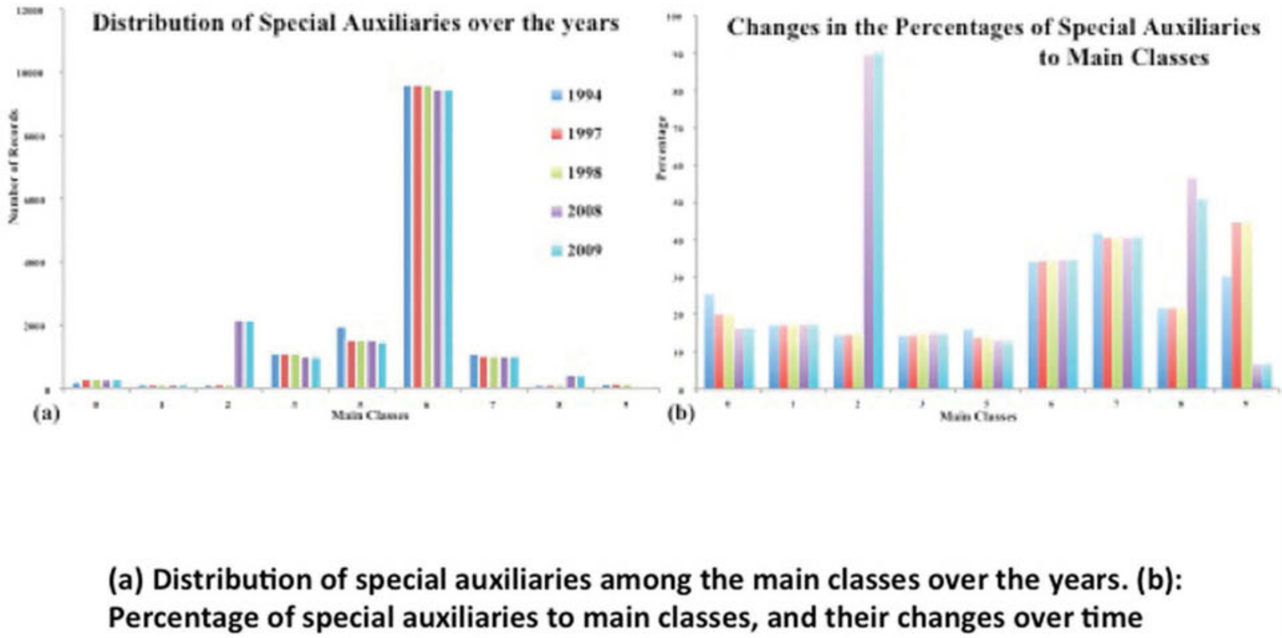

Figure 3. Change in special auxiliaries in the UDC over time (Salah et al. 2012, 53).

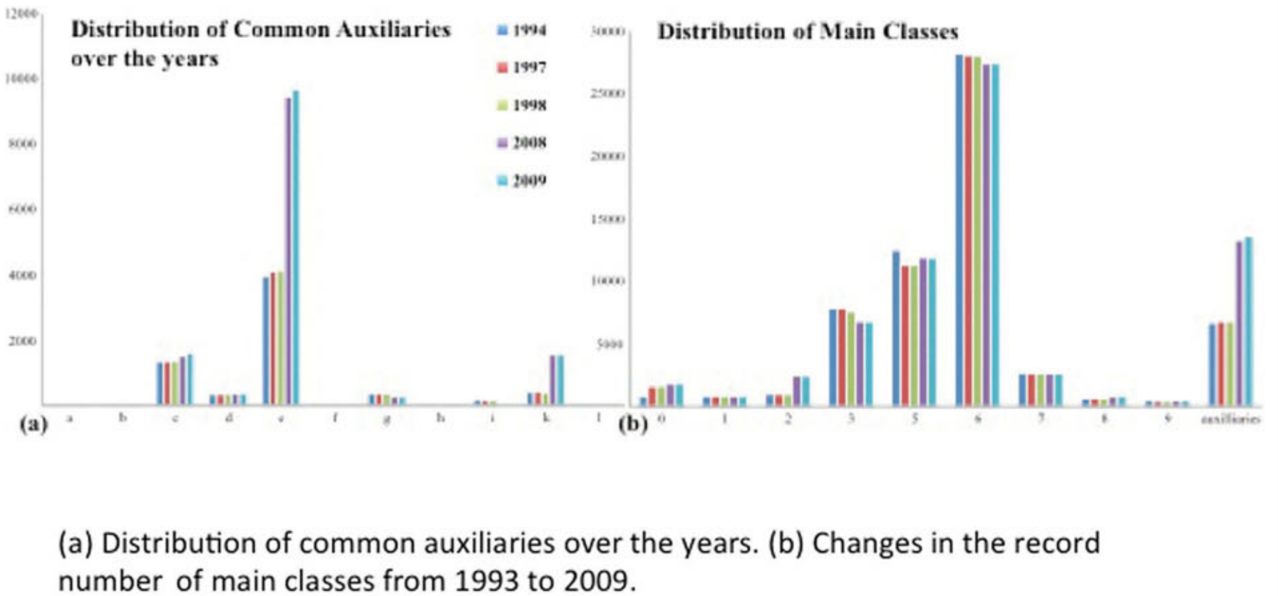

Figure 4. Distribution of common auxiliaries and changes in main classes

the growth of those main classes. Changes in auxiliaries also reflect an editorial practice which aims to preserve UDC numbers in terms of main classes and encapsulates changes using the combinatorial power of the UDC as provided by auxiliaries.

There is another aspect of the UDC that deserves closer inspection. Earlier we mentioned that the network of categories in Wikipedia is far from a tree hierarchy. The classes in the UDC do form such a tree. But the UDC is not designed to pinpoint a concept to a specific place in an otherwise hierarchical system. Its power is the ability to combine simple concepts into more complex and express the interplay of different concepts in a specific string. The main instruments to do so are the common auxiliary signs and their use in the Master Reference File (see Figure 4), which indicates the importance of this element of the UDC considered as a language. This is expressed by the UDC consortium (2016) in the scope note: "The level of detail and specificity of UDC cannot be observed based on the hierarchy levels or from the number of UDC classes as most of compound and complex subjects are described through combination of simple UDC numbers in the process of indexing." Already, the MRF contains a significant number of compound number and consequently the length of a UDC string is regularly larger than six, the maximum length of a single UDC number. Consequently, the application of common auxiliaries turns the UDC tree of classes into a network of concepts (Smiraglia at al. 2013). We will not pursue the presentation of this network characteristic in this article, but the analysis of the complex nature the UDC apparent in its design triggered our curiosity about looking into the application of the UDC in the indexing process, or, if you will, looking at the UDC in the wild, which is what we now call the population of the UDC.

3.0 The population of the UDC

This other phase of our research was devoted to attempts to gain a better understanding of the UDC by analyzing actual UDC usage to ascertain the population of the UDC in different environments. That is, we wanted to know which parts of the whole UDC were actually populated by the assignment of bibliographic entities. We have developed a way to visualize which elements of the UDC were used and to what extent over time. The original KSL team received a file of nine million UDC numbers from the OCLC WorldCat. This would provide a picture of the use of UDC on a global scale, but also we were interested in how it had been used in a particular library. We were able to acquire a complete set of UDC numbers from the online catalog of the library of the Katholische Universiteit Leuven. These were analyzed and that analysis was reported in Smiraglia et al. (2013). In 2015, we received three more data sets, this time from Portuguese sources: the BNP PORBASE "Base Nacional de Dados Bibliográficos," the BNP Catalogo catalog of the National Library of Portugal, and the BND Livre, the National Digital Library of Portugal (Biblioteca Nacional Digital).

3.1 Datasets

From the OCLC WorldCat we received in January 2013, a matched set of 9,055,623 OCLC record numbers and UDC strings from USMARC field 080. Removing pairs with blank 080 fields, and those that carried identifiable non-UDC strings left 8,374,040 pairs. Each pair represents a UDC number assigned to a resource represented by the data in the OCLC USMARC record. They are not necessarily individual resources, as several UDC strings often are assigned to the same resource. We analyzed the UDC strings without regard to the resources to which they were assigned. At the same time, from the LIBIS online catalog of the libraries at KU Leuven we received 95,544 local MARC strings in field $$8, which typically contains both a UDC string and a text string derived from the UDC schedules. In this case, we aggregated unique occurrences, leaving a total of 91,132 UDC strings for analysis. In March 2015, we received datasets of bibliographic records from three Portuguese national resources, including assigned UDC strings. The Portuguese sources promised to provide various approaches both to verify earlier observations and also to diversify results. A primary consideration was that Portuguese libraries use UDC as a form of subject indexing (as does KU Leuven but not most libraries contributing to the WorldCat) rather than for shelving. Figure 5, for example, shows a record from PORBASE with multiple UDC strings assigned.

Thus, we could expect more and more complex UDC strings in the Leuven and Portuguese files than we encountered generally in the WorldCat. Also, as results demonstrate, sources and dates of publication vary regionally, and we wondered whether that would have any visible effect on the population of the UDC. For all three Portuguese collections, we received files created using Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), which yielded XML files containing local records in UNIMARC format. We were able to use the MARC field coding to extract specific data. From BNP PORBASE, we received 1.1 million records of which 530,412 had coded dates of publication and 349,029 had more than one UDC string. From the BNP Catalogo, we received approximately 880,000 records of which 338,505 had usable dates of publication and 369,718 had more than one UDC class. From the BND Livre, we received 21,000 records of which 2670 had usable dates of publication and 12,437 had more than one UDC string.

The three Portuguese datasets are not independent from each other. PORBASE stands for "Base Nacional de Dados Bibliográficos." This catalogue was founded in 1886, coordinated by the National Library, and has been available to the public online since 1988. PORBASE is not only the collective national online catalog, it is also an organization that has set a number of standards, and to be included in it requires an application process. One of the principles is to use the classification via UDC as a means of harmonizing among the different ways works coming from other collections are indexed. The catalog of the National Library (BNP Catalogo) allows seamless searching through all of their collections. BND Livre is the National Digital Library of Portugal (thus, Biblioteca Nacional Digital). BND allows access to digital and digitized content. At present, it has about 25,000 titles including books, periodicals, iconography, cartography, and music and is also a partner with the Europeana digital li-

```
<record><header><identifier>oai:oai.bn.pt:porbase/13</identifier>
<datestamp>2014-11-16</datestamp></header><metadata>
<mx:record xmlns:mx="info:lc/xmlns/marcxchange-v1" format="Unimarc" type="bibliographic">
<mx:leader>00798cam 02200253 04500</mx:leader><mx:controlfield tag="001">13</mx:controlfield><mx:controlfield
tag="005">20010430130900.0</mx:controlfield>
<mx:datafield tag="035" ind1=" " ind2=" "><mx:subfield code="a">(bn)13</mx:subfield></mx:datafield>
<mx:datafield tag="095" ind1=" " ind2=" "><mx:subfield code="a">PTBN00000014</mx:subfield></mx:datafield>
<mx:datafield tag="100" ind1=" " ind2=" "><mx:subfield code="a">19790701d1978    m  y0pora0103
ba</mx:subfield></mx:datafield>
<mx:datafield tag="101" ind1="1" ind2=" "><mx:subfield code="a">por</mx:subfield><mx:subfield
code="c">eng</mx:subfield></mx:datafield>
<mx:datafield tag="102" ind1=" " ind2=" "><mx:subfield code="a">PT</mx:subfield></mx:datafield>
<mx:datafield tag="200" ind1="1" ind2=" "><mx:subfield code="a">&lt;A &gt;casa da
esquina</mx:subfield><mx:subfield code="f">Enid Blyton</mx:subfield><mx:subfield code="g">trad. Maria Helena
Mendes</mx:subfield></mx:datafield>
<mx:datafield tag="210" ind1=" " ind2=" "><mx:subfield code="a">Lisboa</mx:subfield><mx:subfield
code="c">Meridiano</mx:subfield><mx:subfield code="d">1978</mx:subfield></mx:datafield>
<mx:datafield tag="215" ind1=" " ind2=" "><mx:subfield code="a">197, [2] p.</mx:subfield><mx:subfield
code="c">il.</mx:subfield><mx:subfield code="d">19 cm</mx:subfield></mx:datafield>
<mx:datafield tag="304" ind1=" " ind2=" "><mx:subfield code="a">Tít. orig. : House at the
corner</mx:subfield></mx:datafield>
<mx:datafield tag="675" ind1=" " ind2=" "><mx:subfield code="a">087.5</mx:subfield><mx:subfield
code="v">BN</mx:subfield><mx:subfield code="z">por</mx:subfield><mx:subfield
code="3">287999</mx:subfield></mx:datafield>
<mx:datafield tag="675" ind1=" " ind2=" "><mx:subfield code="a">821.111-93"19"</mx:subfield><mx:subfield
code="v">BN</mx:subfield><mx:subfield code="z">por</mx:subfield><mx:subfield
code="3">590415</mx:subfield></mx:datafield>
<mx:datafield tag="675" ind1=" " ind2=" "><mx:subfield code="a">821.111-311.3"19"</mx:subfield><mx:subfield
code="v">BN</mx:subfield><mx:subfield code="z">por</mx:subfield><mx:subfield
code="3">539933</mx:subfield></mx:datafield>
<mx:datafield tag="700" ind1=" " ind2="1"><mx:subfield code="a">Blyton</mx:subfield><mx:subfield
code="b">Enid</mx:subfield><mx:subfield code="f">1897-1968</mx:subfield><mx:subfield
code="3">14734</mx:subfield></mx:datafield>
<mx:datafield tag="702" ind1=" " ind2="1"><mx:subfield code="a">Mendes</mx:subfield><mx:subfield code="b">Maria
Helena</mx:subfield><mx:subfield code="4">730</mx:subfield><mx:subfield
code="3">76937</mx:subfield></mx:datafield>
<mx:datafield tag="801" ind1=" " ind2="0"><mx:subfield code="a">PT</mx:subfield><mx:subfield
code="b">BN</mx:subfield><mx:subfield code="g">RPC</mx:subfield></mx:datafield>
<mx:datafield tag="966" ind1=" " ind2=" "><mx:subfield code="l">BN</mx:subfield><mx:subfield
code="m">FGMON</mx:subfield><mx:subfield code="s">P. 8131 P.</mx:subfield></mx:datafield>
<mx:datafield tag="998" ind1=" " ind2=" "><mx:subfield code="a">FSE01 – 00014</mx:subfield></mx:datafield>
</mx:record></metadata></record>
```


Figure 5. PORBASE bibliographic record with multiple UDC strings.

brary, and its content is harvested by the European Library. For this first analysis, we treated them as separate datasets, but there certainly is an overlap among them, given that PORBASE is the union catalogue and that BND is a collection of specific works from the BNP.

A note on how we processed UDC strings is in order. As noted above, the UDC is unique among bibliographic classifications in its synthetic flexibility and its hospitality to faceted expression. That means, one is not forced into a simple collocating decision about how to assign a text to a large class. Rather, UDC allows a non-linguistic expression of complex context-dependent content descriptors. This feature is one reason the UDC is so amenable to the research reported here—it is not just a device for grouping books for browsing, rather it is a sophisticated means of parsing the precise content of a resource at a depth level of indexing and expressing those parsed concepts in precise strings. We do not put all documents with reference to cats under "cats." Rather, with UDC, we can, for example, describe domestic long-haired cats in Danish literature for children written in the 20th century.

Earlier, we gave an example of a complex UDC string. In reality, such complex strings are rare, although strings with four or five distinct components are not unusual.

Main class numbers may be appended to each other with a plus sign "316.4+100," a slash "316.4/100," a colon "316.4:100," a double colon "316.4::100," or square brackets "316.4[100]." The meanings are subtly different and usage depends on local custom. In some libraries only one technique or the other is employed while in others all may be used at once. In general, a plus sign means "and," but a slash, a colon, or brackets mean what is called a phase relation or "A (treated in) B." In particular the square brackets introduce a sub-arrangement. 316.4 is the classification for social processes and 100 for philosophy. With the plus sign we have social processes and philosophy; with the other connectors we have social processes from a philosophical perspective. 316.4[100] would indicate a sub-arrangement of social processes in which philosophy forms a distinct division. Readers now should consider the meaning of the opposite expression to understand the unique quality of the UDC. 100+316.4 would be philosophy and social processes (the question then arises, is the plus sign in UDC commutative?); 100:316.4 would mean philosophy from the perspective of social processes. So, in every case, the first symbol identifies the primary domain.

The point is, for processing main classes, we counted the first digit in every string and also any first digit after a plus sign, a colon, or a parenthesis. We did not count any that occurred later in strings. Auxiliaries are introduced with other symbols, and for analysis of network structures within the classification, we constructed matrices of main classes and auxiliaries by counting both the main class and any first numeral after an auxiliary indicator. For example, 15(091) is Psychology (History of). So, in any string with a connector sign, we counted the first symbol in each portion (316.4[100] would get a tick in 3 and 1). And in any string with auxiliaries, we counted the first symbol of each portion; (15(091) would get a tick in main class 1 and a tick in auxiliary 09).

Finally, random samples of each dataset were selected to support another study not described here (see Smiraglia 2013, 2104a, 2014b for sampling details). The samples were drawn at 95% confidence with a projected confidence interval of ±5%, which was calculated to require samples of 329 records each. In all cases, 400 records were drawn into the samples and after deduplication, the samples ranged from 359 to 401 records. In all cases, complete MARC records were identified using the UDC strings and other record identifiers in the original datasets. The study for which these samples were drawn used dates of publication and UDC main class population to demonstrate the accuracy of the samples (all matched the population figures from the earlier stages of research). In the narrative that follows, it is noted when sample data have been used to generate visualizations.

4.0 Results

4.1 Dates of publication

We analyzed dates of publication of the works classified in the manner in which a social scientist might gather demographic data. That is, by learning about the works classified, we can learn something about how the UDC has been populated. The results were mixed, which was interesting. Frequency distributions of the number of works per year of publication for each dataset as represented in the sample data are shown in Figure 6.

Depending on how the figures are reproduced it might be possible to see the details, however, the visual impression of large spikes to the right of each figure is important—it shows us that most of the works classified are published after about 1970. At first, we thought this was an artifact of the OCLC WorldCat and of retrospective conversion of card catalogs, and indeed the Leuven distribution seemed also to follow this trend. But the Portuguese data are quite diverse and caused us to reconsider the situation. In particular, the BNDLivre digital catalog shows almost a flat distribution of dates of publication with odd spikes in 1649 and 1849. It seems that we are looking at the distinct collection characteristics of each collection. If we reconsider the spike in the Leuven collection, it seems to correspond to a collection growth pattern beginning about 1974. PORBASE, like the WorldCat, because it is a national bibliographic utility, has a flatter distribution with the bulk of the dates following 1977, which is roughly the same pattern we saw in the WorldCat. On the other hand, the BNP catalog has a fairly consistent distribution from 1945 until 1995 and then a spike following, which suggests a good representation of works collected by the National Library of Portugal. One way to ground these results would be to collect collection development statistics from each of these libraries and bibliographic utilities, a step we admit we have not yet been able to incorporate.

4.2 Distribution of UDC numbers across classes

Of course we also are interested in the specific population of the UDC. The first point of analysis is the population of the main classes. Specifically, we ask the data which classes of the UDC have works assigned to them and to what extent? In the earliest phase of the study, we compared the UDC strings from the OCLC WorldCat to those from the KU Leuven catalog. In the latest phase of the study, we developed visualizations of the populous classes in the three Portuguese datasets. We use both doughnut and spider visualizations to show all these collections in comparison (Figure 7). The visualization in a

1

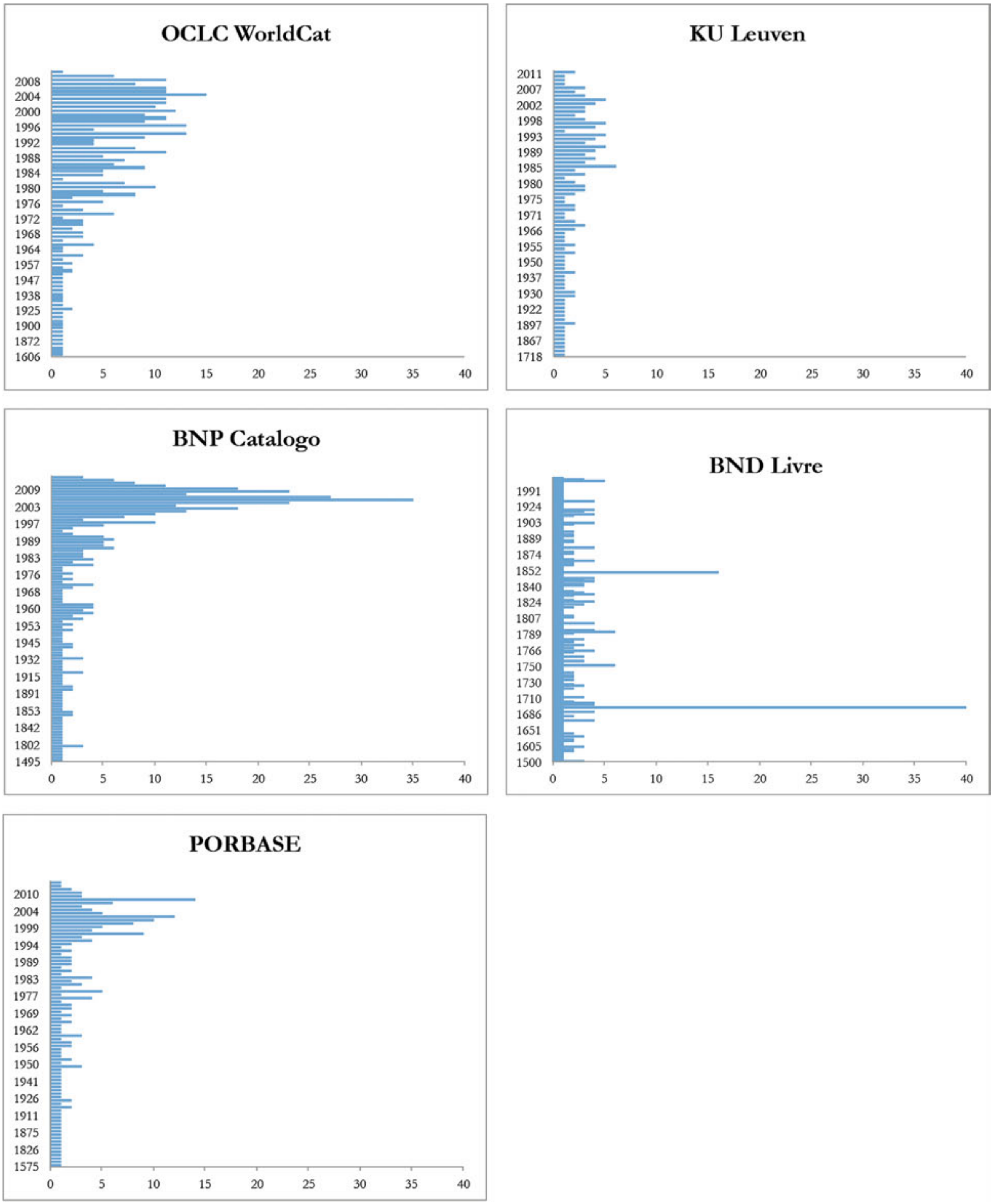

Figure 6. Date of publication of works classified.

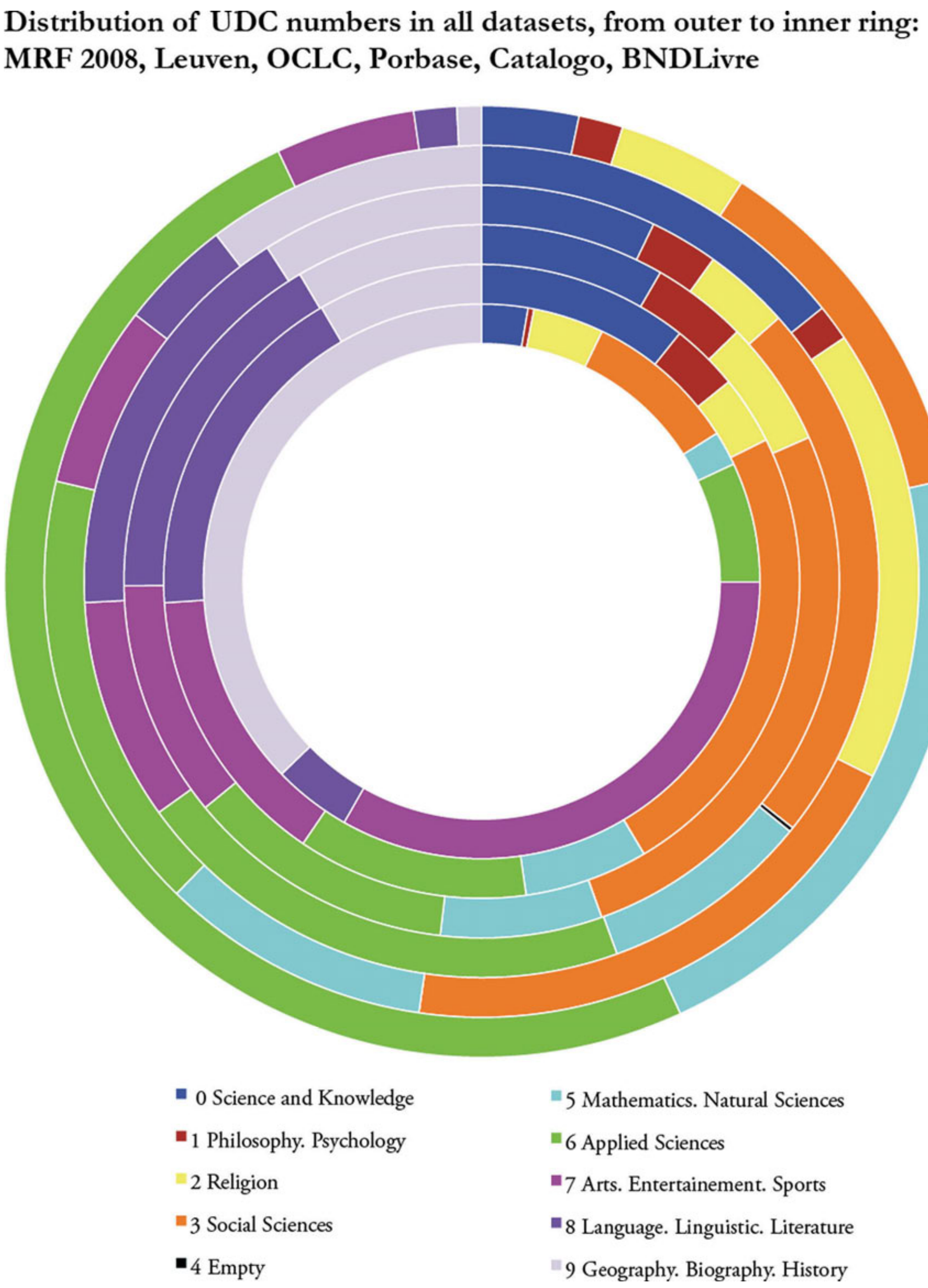

Figure 7. Population of the UDC in all datasets.

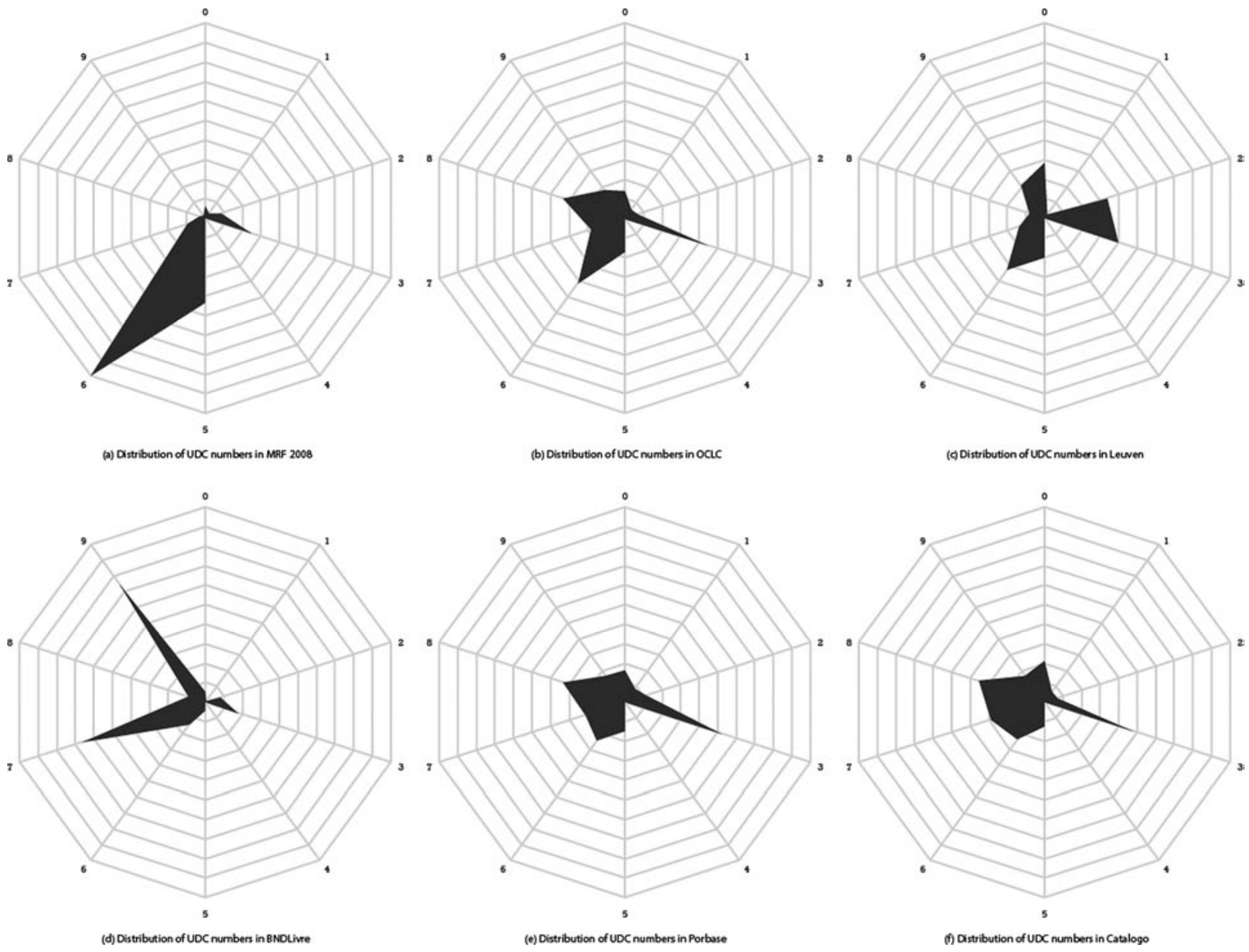

Figure 8. Population of the UDC all datasets:
(a) MRF 2008, (b) OCLC, (c) Leuven, (d) BNDLivre, (e) Porbase, (f) BNP Catalogo.

doughnut gives the relative distribution of the classes, the visualization also shows the size of the different classes, hence the absolute distribution.

The WorldCat distribution is flatter because it represents mixed contributions from many kinds of institutions. All main classes occurred, but the largest clusters are social sciences, applied sciences, and literature. The Leuven distribution has applied sciences, social sciences, and religion as the largest clusters with smaller clusters in arts, history, and natural sciences; little literature or philosophy. The PORBASE distribution shows mostly social sciences with some smattering of the other classes but almost no philosophy or religion. Distributions from BND and BNP are comprised of predominantly history and arts, both with small clusters of social sciences. BNP uses all classes; BND does not.

5.0 Conclusions: functions of visual explorations and knowledge maps

In the beginning, we spoke of four functions that knowledge maps or visual explorations can have. What we have seen now in greater detail is how statistical analysis and visualization can be used to understand better the impact of classification. Throughout this paper, we have demonstrated the power of empirical analysis of the UDC itself, as well as the diverse population of it in different datasets. We have seen that the growth of the UDC over the twentieth century parallels the evolution of knowledge in the academic canon. We have seen that rather than reconstruct main classes with potentially catastrophic revisions, the editors of the UDC have preferred complex and ever more granular evolution of special auxiliaries. And in evaluating the population of the UDC, we have seen even more evidence of the cultural evolution of knowledge across time.

Empirical analysis supports questions from the science of science. How are the works a community relies upon distributed in a disciplinary space? How can subject headings, UDC numbers, and other forms of KOSs be used to determine how far the roots for a certain research topic spread out, and therefore, what kind of reading one needs to recommend to students? With Ginda and Börner we have engaged such an analysis around the history of science dynamics itself, starting with a bibliography as a sample. Smiraglia (2013, 2014a, 2014b) used deconstructed elements of UDC strings to demonstrate their correlation with bibliographic aspects of a collection such as data, place of publication, and language and publisher, among others.

The UDC, both as a curated complex language to express concepts across languages, space, and time as well as a KOS applied by expert cataloguers, invites further analysis. In particular, its network character still hides secrets waiting to be unraveled. How can we interpret the connection between some classes by some common auxiliaries? Do compound UDC numbers, in actual application as well as in design, represent bridges between fields and disciplines, travelling across concepts as we also see in citation links in large scale science maps (see for example Klavans and Boyack 2009)?

In this paper, we discussed how using the UDC can help us to shed light on the evolution and composition of collections. We also showed that the UDC as a reference system has changed in composition over time itself. The deeper and more granular our analysis becomes, the more we need to take into account that most of the UDC in use does not come with version numbers of the Master Reference File used. A forensic analysis of UDC use in comparison to the UDC design process might shed light on the noise we have to take into account when analyzing UDC use combining different editions. For future application of the UDC in knowledge graphs that are machine readable for the semantic web, keeping traces of the UDC's provenance becomes a must. For the retrospective analysis of the UDC or other classifications in collection use, there is still a wide territory to be explored before provenance can become a priority.

KOSs such as the UDC allow us in principle to gain overview about the content of collections, their main disciplinary orientations, their roots in history, and their richness in terms of interwoven concepts in the works they carry. Parsing automatic traces from bibliographic records is a careful business, which in turn, helps to better curate the records themselves. Visualizations, knowledge maps as shown in this paper, are of analytic use but can also be enhanced toward "generous interfaces" (Whitelaw 2015), showcasing the treasures in a collection but also complementing browsing through collections (Mutschke, May, and Scharnhorst 2014).

References

Börner, Katy. 2010. *Atlas of Science: Visualizing What we Know.* Cambridge, Mass.: MIT Press.

Börner, Katy. 2011. "Plug-and-play Macroscopes." *Communications of the ACM* 54 no. 3: 60-69. doi:10.1145/1897852.1897871

Börner, Katy. 2015. *Atlas of Knowledge: Anyone can Map.* Cambridge, Mass.: MIT Press.

Bowker, Geoffrey C. and Susan Leigh Star. *Sorting Things Out: Classification and its Consequences*. Cambridge, Mass: MIT Press.

Ereshefsky, Marc. 2001. *The Poverty of the Linnaean Hierarchy: A Philosophical Study of Biological Taxonomy*. Cambridge: Cambridge University Press.

Ginda, Michael, Andrea Scharnhorst and Katy Börner. 2016. "Modelling the Structure and Dynamics of Science Using Books." In *Theories of Informetrics and Scholarly Communication: A Festschrift in Honor of Blaise Cronin*, ed. Cassidy Sugimoto. Munich: De Gruyter Saur, 304-34.

Heaney, Michael, ed 2009. *Library Statistics for the Twenty-first Century World: Proceedings of the Conference held in Montréal on 18-19 August 2008 Reporting on the Global Library Statistics Project.* München: K.G. Saur.

Kedrov, B. M. 1975-76. *Klassifizierung der Wissenschaften.* Ins Deutsche übersetz durch Lili Keith und L. Pudenkowa. Berlin: Akademie-Verlag.

Klavans, R., & Boyack, K. W. (2009). "Toward a Consensus Map of Science." *Journal of the American Society for Information Science and Technology* 60: 455–76. doi:10.1002/asi.20991

McIlwaine, I. C. 2007. *The Universal Decimal Classification: A Guide to its Use.* Rev. ed. The Hague: UDC Consortium.

Meroño-Peñuela, A. 2016. "Measuring Quality of Evolution in Diachronic Web Schemas Using Inferred Optimal Change Models." Unpublished paper.

Mutschke, Peter, Philip Mayr and Andrea Scharnhorst, eds. 2014. *KMIR 2014—Knowledge Maps and Information Retrieval: Proceedings of the First Workshop on Knowledge Maps and Information Retrieval co-located with International Conference on Digital Libraries 2014—ACM/IEEE Joint Conference on Digital Libraries (JCDL 2014).* CEUR-WS.org.

Noy, Natalya F. and Michael Klein. 2004. "Ontology Evolution: Not the Same as Schema Evolution." *Knowledge and Information Systems* 6: 428-40. doi:10.1007/s10115-003-0137-2

Salah, Almila Akdag, Cheng Gao, Krzysztof Suchecki, Andrea Scharnhorst, and Richard P. Smiraglia. (2012). "The Evolution of Classification Systems: Ontogeny of the UDC." In *Categories, Contexts and Relations in Knowledge Organization: Proceedings of the Twelfth International ISKO Conference 6-9 August 2012 Mysore, India*, ed. A. Neelameghan and K.S. Raghavan. Würzburg: Ergon Verlag, 51–57.

Salah, Almila Akdag, Lev Manovich, L., Alberet Ali Salah and Jay Chow. 2013. "Combining Cultural Analytics and Networks Analysis: Studying a Social Network Site with User-Generated Content." *Journal of Broadcasting & Electronic Media 57*: 409–26. doi:10.1080/08838151.2013.816710

Scharnhorst, Andrea. 2015. "Walking Through a Library Remotely: Why we Need Maps for Collections and How KnoweScape can Help us to Make Them." *Les Cahiers Du Numérique* 11: 103–27. doi:10.3166/lcn.11.1.103-127

Scharnhorst, Andrea and Richard P. Smiraglia. 2012. "Evolution of Classification Systems." *Advances In Classification Research Online* 23: 56. doi:10.7152/acro.v23i1.14264 Scharnhorst, A., C. Gao, A. Akdag Salah and K. Suchecki. 2012. "Evolution of Wikipedia Categories." DANS. http://dx.doi.org/10.17026/dans-xjp-zfuw

Smiraglia, Richard P. 2013. "Big Classification: Using the Empirical Power of Classification Interaction." In *Proceedings of the ASIST SIG/CR Classification Workshop, Montréal, 2 November 2013*, ed. D. Grant Campbell, 21-29. doi:10.7152/acro.v24i1.14673

Smiraglia, Richard P. 2014a. "Classification Interaction Demonstrated Empirically." In *Knowledge organization in the 21st century: Between Historical Patterns and Future Prospects, Proceedings of the 13th International ISKO Conference, Krakow, Poland, May 19-22, 2014,* ed. Wiesław Babik. Advances in Knowledge Organization v. 14. Würzburg: Ergon-Verlag, 176-83.

Smiraglia, Richard P. 2014b. "Extending the Visualization of Classification Interaction with Semantic Associations." In *Proceedings of the ASIST SIG/CR Classification Workshop, Seattle 1 November 2014*.

Smiraglia, Richard P., Andrea Scharnhorst, Almila Akdag Salah and Cheng Gao. 2013. "UDC in Action." In *Classification and Visualization: Interfaces to Knowledge, Proceedings of the International UDC Seminar, 24-25 October 2013, The Hague, The Netherlands,* ed. Aïda Slavic, Almila Akdag Slah and Sylvie Davies. Würzburg: Ergon-Verlag, 259-72.

Suchecki, Krzysztof, Alkim Almila Akdag Salah, Cheng Gao and Andrea Scharnhorst. 2012. "Evolution of Wikipedia's Category Structure." *Advances in Complex Systems* 15 supp01: 1250068.

Tennis, Joseph T. 2012. "The Strange Case of Eugenics: A Subject's Ontogeny in a Long-lived Classification Scheme and the Question of Collocative Integrity." *Journal of the American Society for Information Science and Technology* 63: 1350–59. doi:10.1002/asi.22686

Tennis, Joseph T., Katherine Thornton and Andrew Filer. 2012. "Some Temporal Aspects of Indexing and Classification." In *iConference '12: Proceedings of the 2012 iConference.* New York: ACM, 311–16. doi:10.1145/2132176.2132216

UDC Consortium. 2016. UDC Scope. http://www.udcc.org/index.php/site/page?view=about_scope

Whitelaw, Mitchell. 2015. "Generous Interfaces for Digital Cultural Collections." *DHQ: Digital Humanities Quarterly* 9: 1-16. http://www.digitalhumanities.org/dhq/vol/9/1/000205/000205.html

Wikipedia. 2015. "Category: Main Topic Classifications." https://en.wikipedia.org/wiki/Category:Main_topic_classifications