Toward Better Interoperability of the NARCIS Classification

Gerard Coen*, Richard P. Smiraglia**

*Data Archiving and Networked Services DANS, Anna van Saksenlaan 51, 2593 HW Den Haag, Netherlands, <Gerard.coen@dans.knaw.nl>

**Institute for Knowledge Organization and Structure, Inc., Shorewood, Wisconsin 53211, USA, <Richard.smiraglia@knoworg.org>

Gerard Coen is a Policy Officer at Data Archiving and Networked Services (DANS), where he is also a member of the Research & Innovation Group. Since 2017 he has been working as part of the Knowledge Organization Systems Observatory (KOSo), investigating, how KOSs change over time, and how they can be organized and archived. He also works on topics related to semantic interoperability, the European Open Science Cloud (EOSC) and FAIR data and software.

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.

Coen, Gerard and Richard P. Smiraglia. 2019. "Toward Better Interoperability of the NARCIS Classification." *Knowledge Organization* 46(5): 345-353. 9 references. DOI:10.5771/0943-7444-2019-5-345.

Abstract: Research information can be useful to science stakeholders for discovering, evaluating and planning research activities. In the Netherlands, the institute tasked with the stewardship of national research information is DANS (Data Archiving and Networked Services). DANS is the home of NARCIS, the national portal for research information, which uses a similarly named national research classification. The NARCIS Classification assigns symbols to represent the knowledge-bases of contributing scholars. A recent research stream in knowledge organization known as comparative classification uses two or more classifications experimentally to generate empirical evidence about coverage of conceptual content, population of the classes, and economy of classification. This paper builds on that research in order to further understand the comparative impact of the NARCIS Classification alongside a classification designed specifically for information resources. Our six cases come from the DANS project Knowledge Organization System Observatory (KOSo), which itself is classified using the Information Coding Classification (ICC) created in 1982 by Ingetraut Dahlberg. ICC is considered to have the merits of universality, faceting, and a top-down approach. Results are exploratory, indicating that both classifications provide fairly precise coverage. The inflexibility of the NARCIS Classification makes it difficult to express complex concepts. The meta-ontological, epistemic stance of the ICC is apparent in all aspects of this study. Using the two together in the DANS KOS Observatory will provide users with both clarity of scientific positioning and ontological relativity.

Received: 30 November 2018; Revised: 22 May 2019; Accepted: 29 May 2019

Keywords: classification, NARCIS, NARCIS Classification, Information Coding Classification, ICC, sciences, research information

1.0 Research information and classification

Research information serves many purposes. It can be useful to science stakeholders for discovering, evaluating and planning research activities at both national and regional levels. The principle stakeholders of research information are: researchers, research managers, policy-makers, research councils, the media and the public (Nabavi, Jeffery and Jamali 2016). In the Netherlands, the institute tasked with the stewardship of national research information is DANS (Data Archiving and Networked Services). DANS is an institute of the Royal Netherlands Academy of the Arts and Sciences (Koninklijke Nederlandse Akademie van Wetenschappen or KNAW) and the Netherlands Organisation for Scientific Research (Nederlandse Organisatie voor Wetenschappelijk Onderzoek or NWO). DANS is the home of NARCIS, the national portal for research information. It hosts the Netherlands' wide-ranging data and research archiving structure (https://dans.knaw.nl/en).

The Dutch research infrastructure makes use of a national research classification named after the National Academic Research and Collaborations Information System (NARCIS), a DANS-maintained national research portal. NARCIS is an information repository of (open access) publications and datasets from Dutch scholars, combined with texts of peer reviewed publications and other research data (https://dans.knaw.nl/en/about/services/narcis). The NARCIS Classification is a framework with which institutes and experts (scientists and scholars) are identified and clustered symbolically in the infrastructure from which search and retrieval are facilitated. The classification is designed to provide access to research information (e.g., researcher names, institutional affiliations, etc.). The NARCIS Classification assigns symbols to represent the knowledge-bases of contributing scholars, rather than to represent the content of the publications in its repository.

A recent analysis (Smiraglia 2017) demonstrated the strengths of the NARCIS classification as its disciplinary base and its grounding in the NWO/KNAW milieu. The classification has eight main classes: D10000 Science and Technology, D20000 Life Sciences, Medicine and Health Care, D30000 Humanities, D40000 Law and Administration, D50000 Behavioural and Educational Sciences, D60000 Social Sciences, D70000 Economics and Business Administration and E10000 Interdisciplinary Sciences. The classes are seen as mutually exclusive and collectively exhaustive and are derived from the institutional participants whose data comprise the NARCIS portal.

A recent research stream in knowledge organization known as comparative classification uses two or more classifications experimentally to designate content from a test collection (Szostak and Smiraglia 2017a; 2017b; 2018). The purpose is to generate empirical evidence about: a) coverage of conceptual content, including precision; b) population of the classes; and, c) economy of classification (in other words, how many expressions are required to represent the content fully, comparatively?).

This paper builds on that research in order to further understand the comparative impact of the NARCIS Classification in a particular DANS application alongside a classification designed specifically for information resources. Our sample data come from the DANS project Knowledge Organization System Observatory (KOSo), which itself is classified using the Information Coding Classification (ICC) created in 1982 by Ingetraut Dahlberg, the founder of the International Society for Knowledge Organization and principal proponent of a science of knowledge organization. ICC is considered to have the merits of universality, faceting and a top-down approach.

Our exploratory study has the following research question:

> How do NARCIS Classification and Information Coding Classification compare when applied to knowledge organization systems in the social sciences, humanities and life sciences with regard to:
>
> a) coverage of conceptual content, including precision;
> b) population of the classes; and,
> c) economy of classification?

This research has several implications. First, in making this comparison we can discover similarities and differences in the classification of a set of resources using the NARCIS Classification and the ICC. From an academic point of view, the research points to the potential interoperability of the NARCIS Classification. Practically for DANS, the results can diagnose existing limitations of the NARCIS Classification and potentially highlight the path towards improvement, at the same time incorporating the new Observatory more fully into other DANS data resources.

2.0 Methodology

Our method is straightforward. All entities in the KOSo are classified using the ICC. Therefore, we simply also applied NARCIS classification codes (https://www.narcis.nl/classification/Language/en) to each entity. It is important to understand that the ICC was developed for information environments, which is to say it is a meta-disciplinary classification (Bates 1999; Dahlberg 2008). The ICC works for the KOSo by using together codes from Dahlberg's 1999 Classification System for Knowledge Organization Literature. For example, The Canadian Parks Service Classification is classified as "048" (classification)-9 (geographic reference). For the purpose of this comparative analysis we have used only the ICC codes and not those designating types of systems.

3.0 Results

3.1 Population of the two classifications

The KOSs were classified into twenty-eight cases from the life sciences and 299 from the social sciences and humanities—all those in the KOSo as of August 2018. All but two cases from the life sciences were assigned existing NARCIS classification codes. Figure 1 shows the most populous classes from the ICC coding. Most of the KOSs in the sample fell into biology and agriculture, with fewer in aspects of medicine.

All of the NARCIS classification codes came from the D2xxxx hierarchy; D identifies the "Science and Technology" hierarchy, and D2xxxx identifies "Life Sciences, Medicine and Health Care." Almost half of the NARCIS codes

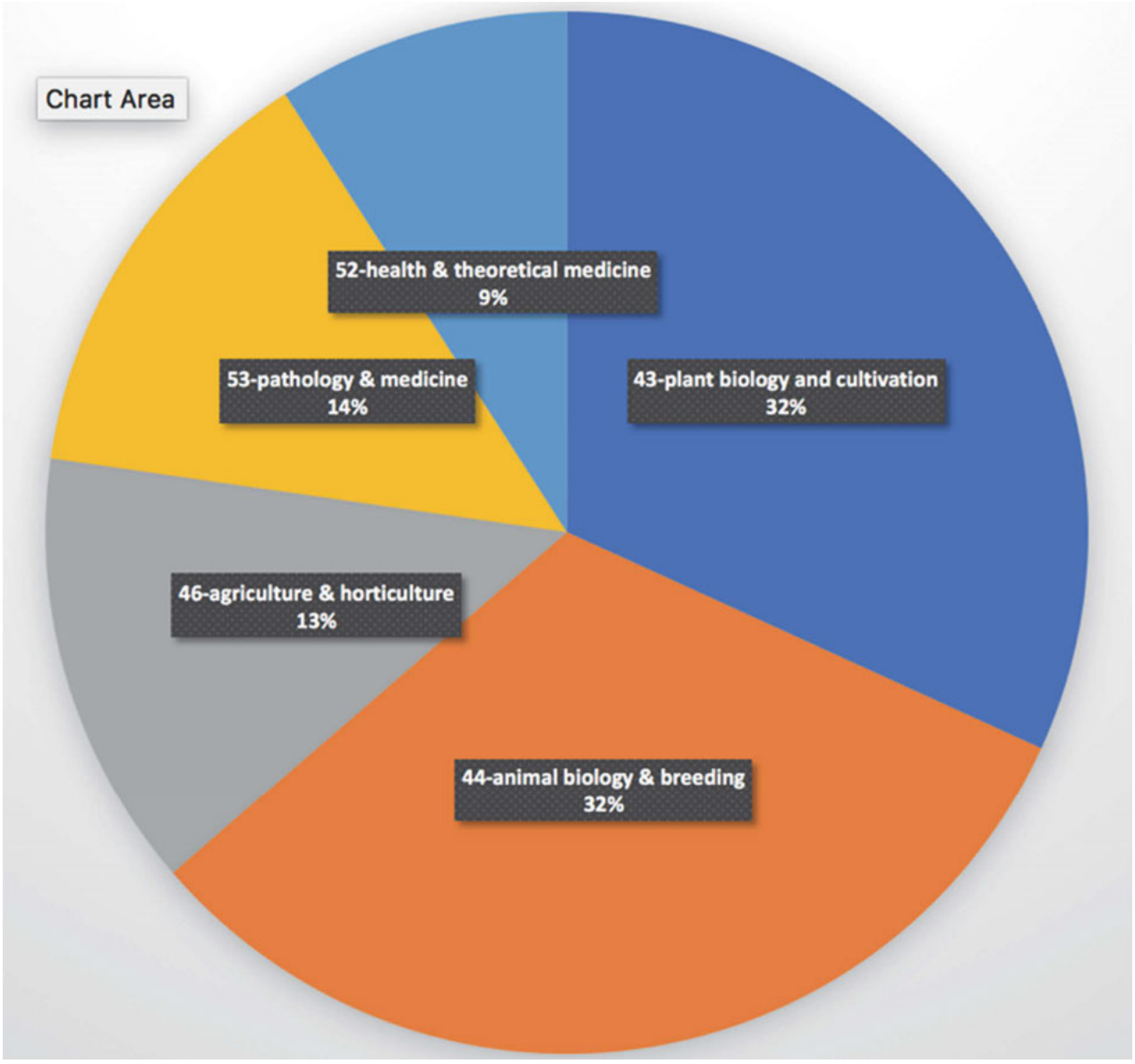

Figure 1. Most populous ICC classes in life sciences KOS sample.

applied fell in class D22600 "Zoology" within the D22000 "Biology" hierarchy. Smaller clusters fell in D21300 "Biochemistry" under "Life Sciences" and D23100 "Pathology" within "Medicine." Thus, in the life sciences, granularity is not different in either classification, although ICC was more flexible in two cases.

There was greater divergence in the social sciences and humanities, where 127 cases could not be classified with NARCIS classification codes, and 138 did not have ICC coding. Fifteen KOSs required two separate NARCIS classification codes to achieve coverage of the concepts. Figure 2 shows the most populous ICC classes in the social sciences and humanities sample. Most works fall into culture or the fine arts.

The majority of the NARCIS codes fall into the D3xxxx "Humanities" range. The distribution of the most populous NARCIS classes is shown in Figure 3. The largest classes are "Paleography – library [sic] science" (for information sciences, especially taken together with computer science), economics and arts and culture.

Once again, although the sample is much more diverse, there is little difference between the two classifications in coverage or granularity.

3.2 Six case examples

As a kind of simple case study, we present six examples of comparative classification using NARCIS and ICC, three each from the social sciences and humanities and from the

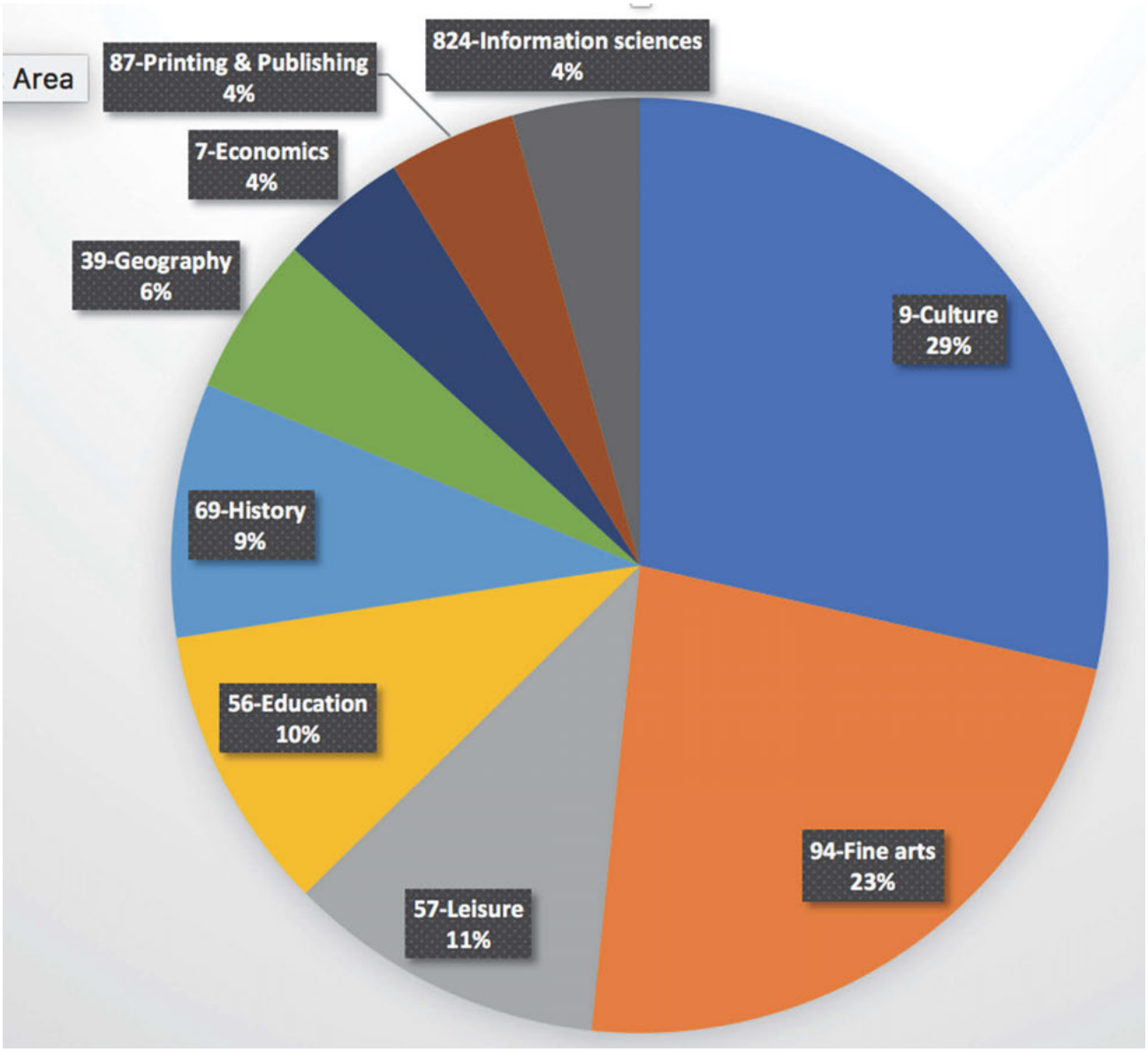

Figure 2. Most populous ICC classes in social sciences and humanities KOS sample.

life sciences. Our first example (Figure 4) uses the Historical International Standard Classification of Occupations (HISCO).

In this case, the ICC has placed all professions within the ontological meta-category of "Human Area," and has broken the large class including profession, labor and leisure into a smaller class including professions and occupations, which presumably include personnel administration. NARCIS has relied solely on economics as a science and personnel administration as a subdivision of that science. The difference in both coverage and granularity in this case is apparent and arises from the narrow scientific focus of NARCIS over and against the more general aspirations of the ICC.

Our next example (Figure 5) is the Europeana Fashion Vocabulary (EFV).

Fashion presents an interesting example of the difficulty of close classification of a domain that lies outside of typical academic interests. There is no good place for "fashion" in either classification. Wikipedia defines "fashion" as style in clothing and life accessories; and the *EFV* itself intends to represent the unity of all meanings of fashion. Using the NARCIS classification we can assign the code for "arts and culture" within the "humanities," but we cannot express the concept of commodity science. The code D14700 "Industrial Design" might be applicable in this case, however, only in combination with another code capable of capturing the social, symbolic and cultural aspects of fashion. In the ICC, we are able to express both "fine

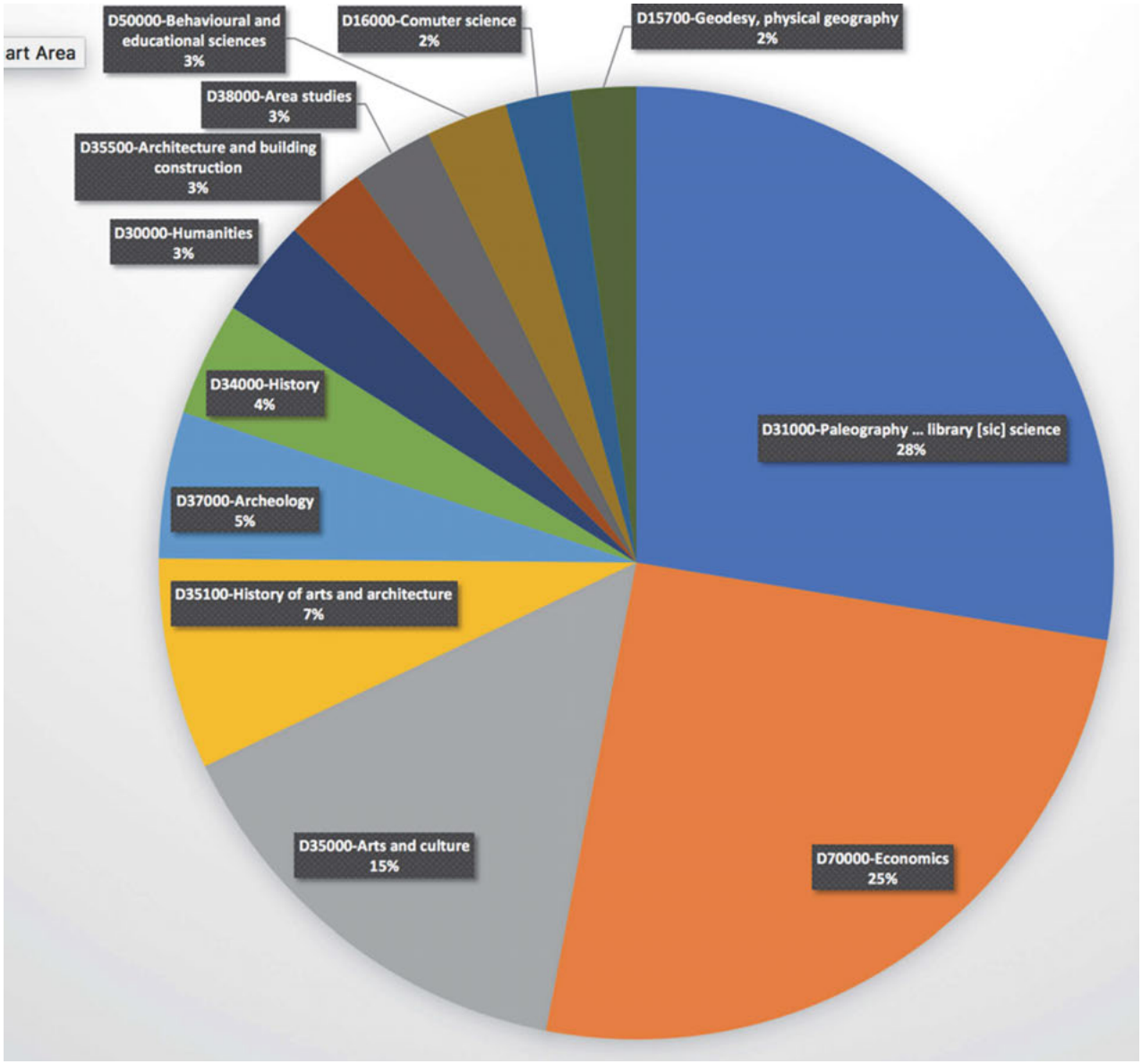

Figure 3. Most populous NARCIS classes in social sciences and humanities KOS sample.

arts" as an aspect of "culture" and "textiles, leather & fur technology" as an aspect of commodity, which is closer yet still imprecise. Using the ICC, it is possible to synthesize codes using "combinatory functions" representing the expression of complex concepts (Dahlberg 2008); Dahlberg suggests using symbols from the Universal Decimal Classification (UDC) to accomplish synthesis. Thus, in the case of the *EFV* we can combine "765 Textiles, leather & fur technology" with "94 Fine arts) as "048-765+940."

The *EFV* case is a good example of the limits of the NARCIS classification, which is derived from the research infrastructure of Dutch universities and KNAW. The NARCIS database returns over 3,500 results for the term "fashion," most of which are not classified. The top results among projects are projects concerning fashion in Dutch society—these are classified as D51000 "Psychology," D35100 "History of arts and architecture" and D63000 "Cultural anthropology." In the UDC, fashion appears at 391 under "Cultural anthropology" and 685.34 "Fashion" and 687.5 "Aesthetics," both of which fall within industry.

Our final social science and humanities example is the well-known ICONCLASS iconographic classification frequently used in art history (Figure 6).

In this case, ICC is able to express "art history" precisely, while NARCIS offers less precision. Interestingly, NARCIS would provide a close fit for the related *Art & Architecture Thesaurus*, because it expresses both art and architecture in one class.

Our first life science example is the International Statistical Classification of Diseases (ICD) (Figure 7).

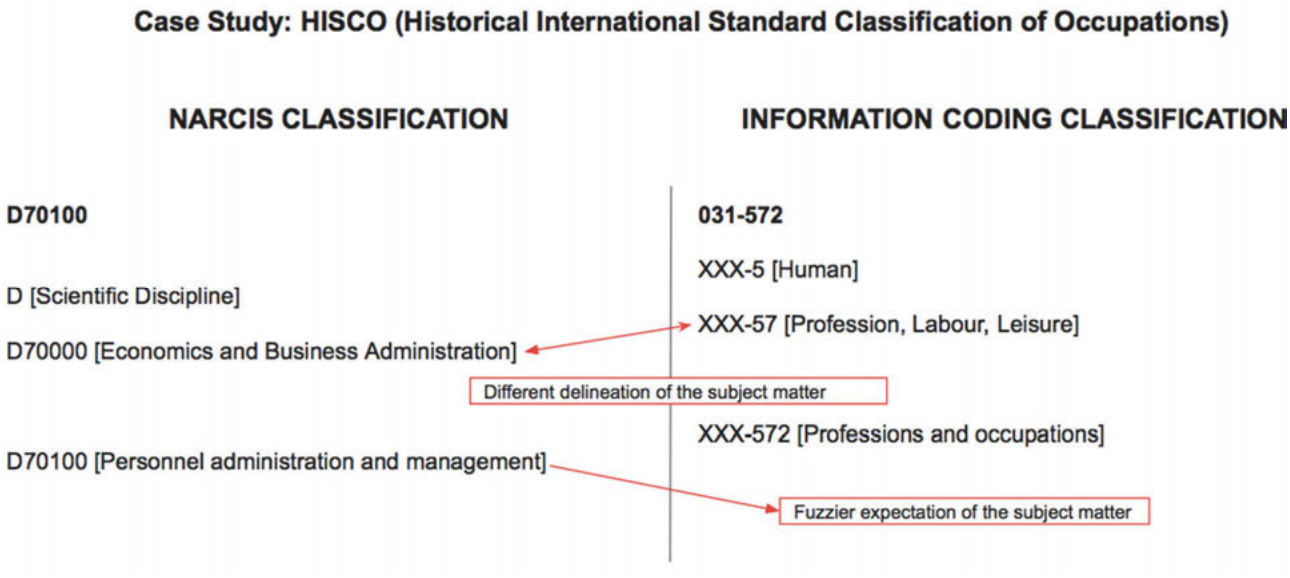

Figure 4. Comparative classification of the Historical International Standard Classification of Occupations.

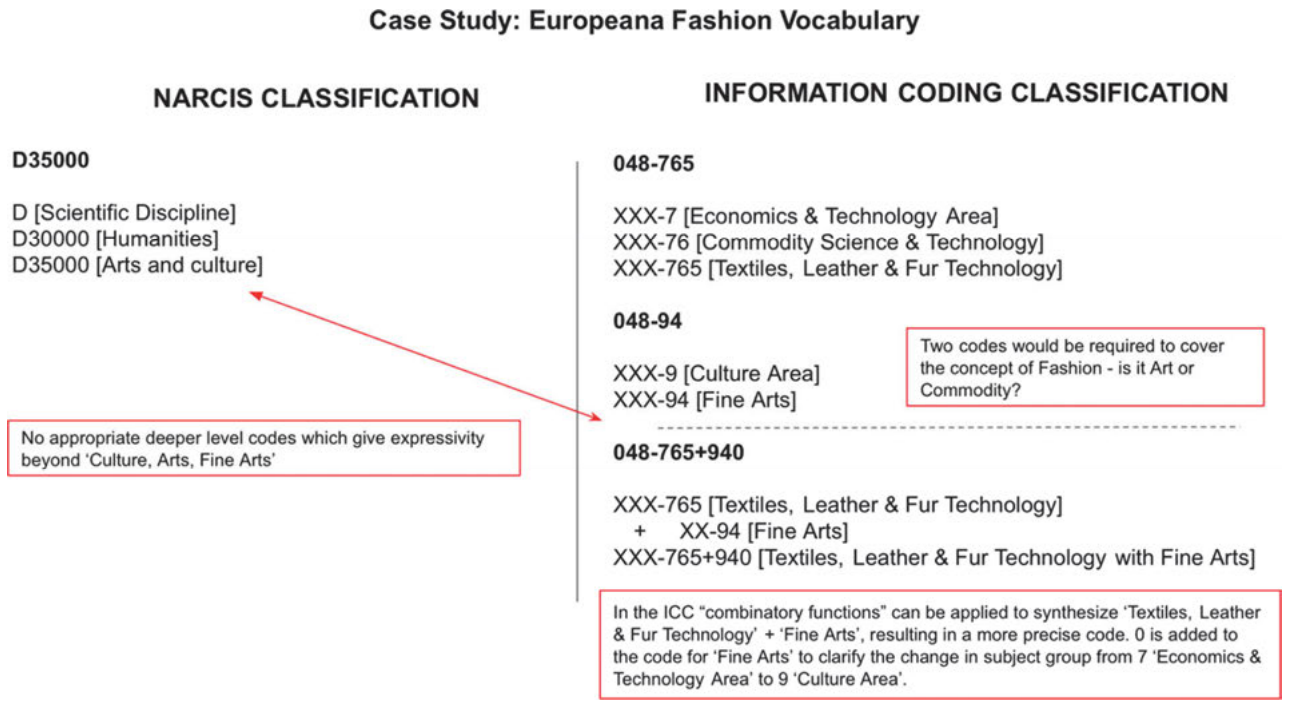

Figure 5. Comparative classification of the Europeana Fashion Vocabulary.

Disease falls within pathology, which falls within medicine in both classifications. The ICC has a broader upper ontological level, however, of "Human Area," incorporating many aspects of human life beyond the life sciences. The obverse is also apparent, that NARCIS is more closely focused directly on the sciences involved. ICC offers some potential for greater granularity than does NARCIS.

Our next example, the Catalogue of Life (CoL), is a species taxonomy that aims to index the world's known species of animals, plants, fungi and micro-organisms. Neither classification reaches the precision to describe the CoL without using two codes, the alternative is leaving the KOS classified under the rather broad bucket of "Biology/Bio-Area" (Figure 8).

Both classifications require multiple class assignments to cover both botany and zoology. The enumerative style of the NARCIS Classification generates each domain in its own class under biology. There is slightly more expressivity

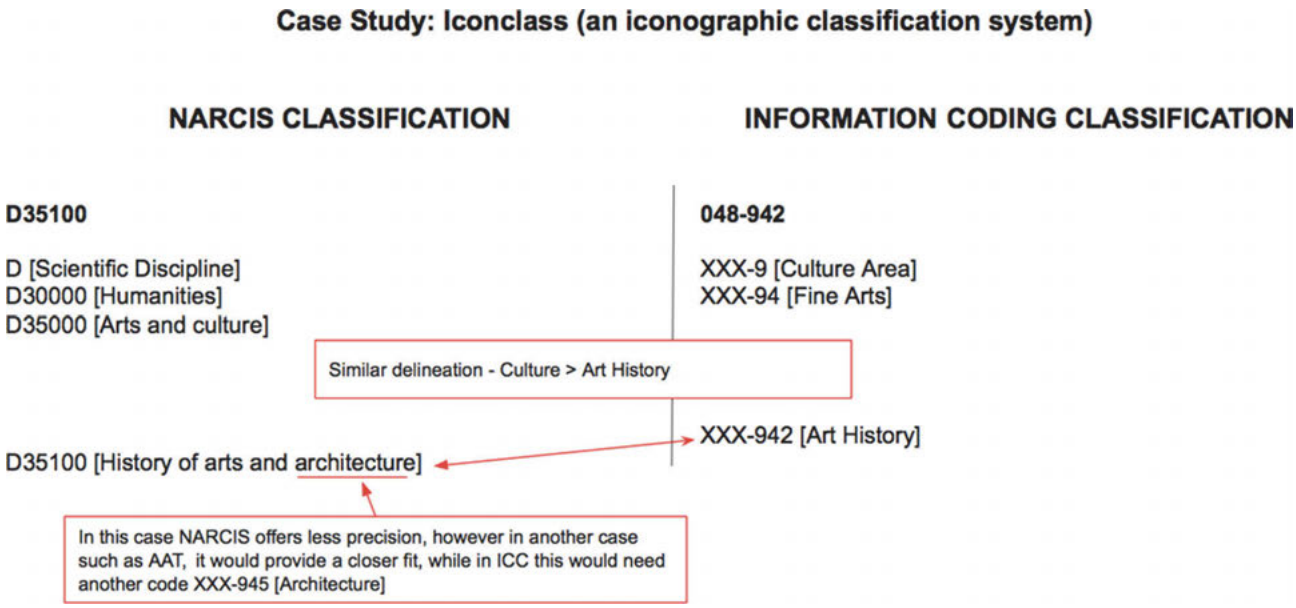

Figure 6. Comparative classification of ICONCLASS.

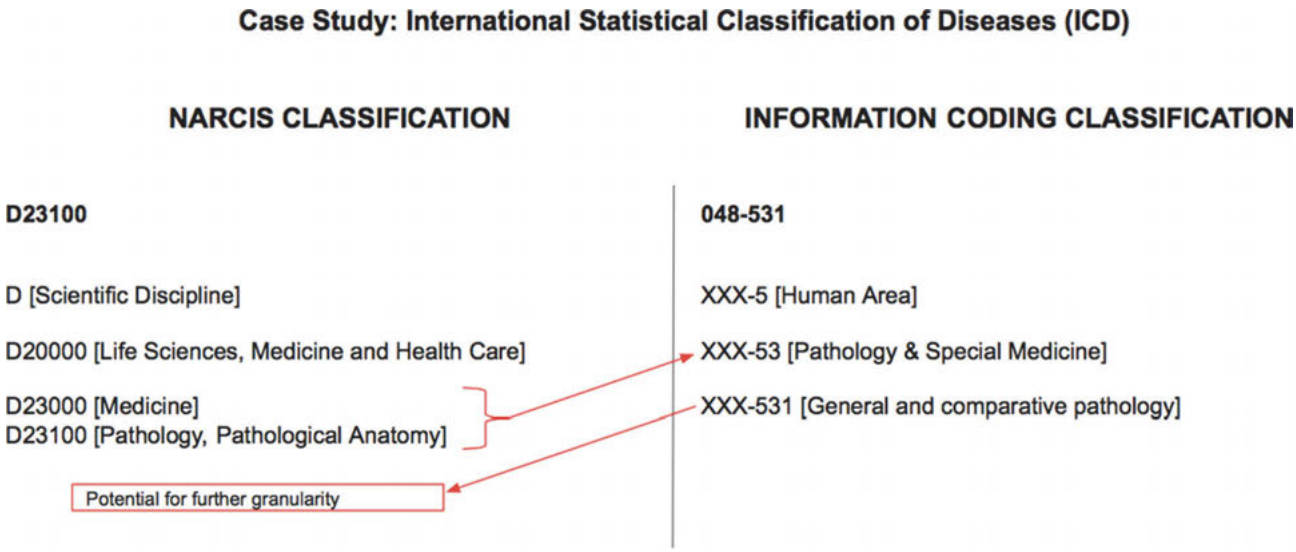

Figure 7. Comparative classification of the International Statistical Classification of Diseases.

in the ICC, where we are able to combine "General botany" with "General zoology" thus: "048-433+43."
Our last example is the Chemical Entities of Biological Interest (ChEBI) ontology of molecular entities (Figure 9).

In this case, both classifications offer a precise location; concepts relating to "hard science" are explicitly expressed in both systems.

4.0 Discussion

We wanted to compare two ontological structures for the subjects of research. The Dahlberg Information Coding Classification was designed specifically for research, and the NARCIS Classification was designed specifically for the research and datasets in the NARCIS database. Thus, the comparison should tell us whether either or both provide appropriate granularity and coextensively—that is, can concepts in research be represented directly and fully in one system or the other, or are the two comparable. We also wanted to see how using the NARCIS Classification alongside ICC might enhance the information infrastructure at DANS. At DANS, the ICC is used as the primary classification for the KOSo project and NARCIS Classification is used for everything else. Would using the two together provide complementarity or redundancy?

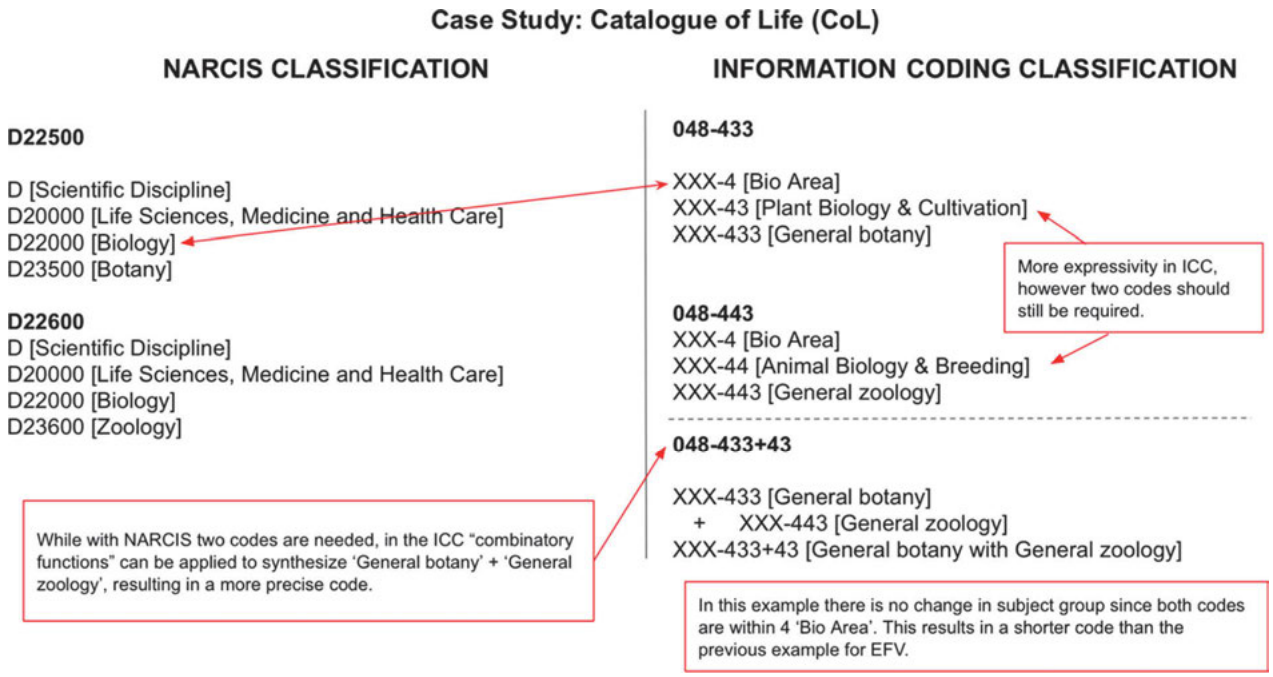

Figure 8. Comparative classification of the Catalogue of Life (CoL).

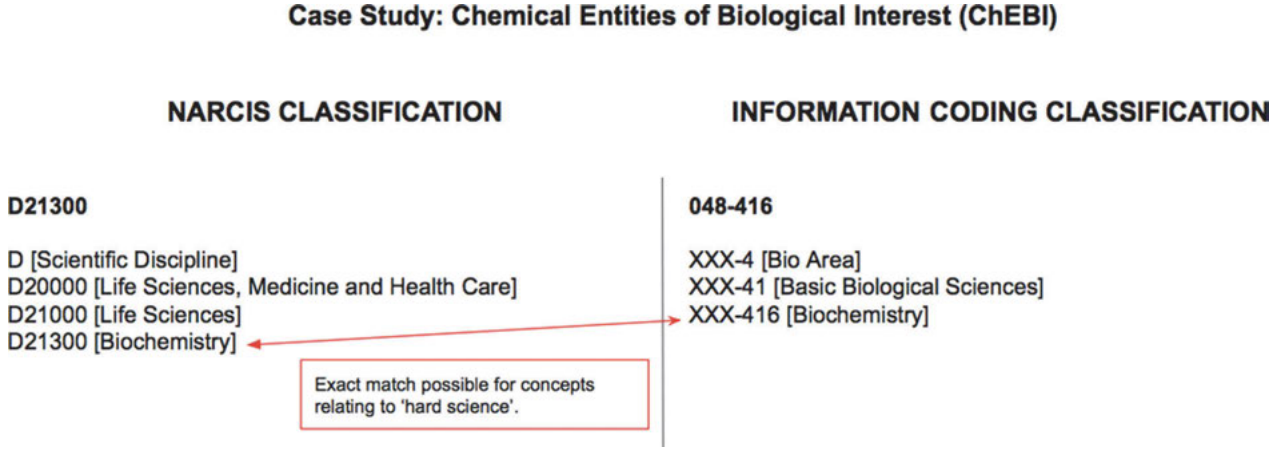

Figure 9. Comparative classification of Chemical Entities of Biological Interest (ChEBI).

In the social science and humanities cases, two—occupations and art history—were provided appropriate granularity in ICC. Fashion was not provided a hospitable location in either. In the life sciences cases, one case—diseases—proved hospitable only at a meta level but not precisely. Botany and zoology required multiple class assignments from both systems to express the relation of the two in one context. In the last—biochemistry—both systems offer precision.

5.0 Conclusions

This study was intended to be exploratory, with the goal of creating better empirical understanding of how these two classifications might contribute to the enrichment of resources in DANS, in particular the KOSs described in the Observatory. Results, therefore, are themselves exploratory. We see that both classifications provide fairly precise coverage. In only a few cases was it difficult to find a precise code in either classification. We see also that the inflexibility of the NARCIS Classification makes it difficult to express complex concepts, such as those expressed more easily in ICC and other general classifications. The meta-ontological epistemic stance of the ICC is apparent in all aspects of this study. That is, NARCIS provides more clarity and granularity in the representation of sciences, but ICC provides better ontological structuring of sciences within other realms of human knowledge. This re-

sult is tautological in part, because of course, those are divergent goals of the two classifications. It does suggest that using the two together in the DANS KOS Observatory will provide users with both clarity of scientific positioning, on the one hand, and ontological relativity, on the other. It would be useful in the application of the NARCIS classification in the Observatory if it were more flexible in terms of offering means of synthesis for expressing complex concepts.

References

Bates, Marcia J. 1999. "The Invisible Substrate of Information Science." *Journal of the American Society for Information Science* 50: 1043-50. DOI: 10.1002/(SICI)1097-4571(1999)50:12<1043::AID-ASI1>3.0.CO;2-X

Dahlberg, Ingetraut. 1999. "Classification System for Knowledge Organization Literature." *Knowledge Organization* 26: 192-202.

Dahlberg, Ingetraut. 2008. "Information Coding Classification: (ICC): A Modern, Theory-Based Fully-Faceted, Universal System of Knowledge Fields" *Axiomathes* 18: 161-76. DOI: 10.1007/s10516-007-9026-8

Fashion. 2018. *Wikipedia.* https://en.wikipedia.org/wiki/Fashion

Majid Nabavi, Keith Jeffery and Hamid R. Jamali. 2016. "Added Value in the Context of Research Information Systems." *Program* 50: 325-39. DOI: 10.1108/PROG-10-2015-0067

Smiraglia, Richard P. "Disciplinary, Asynthetic, Domain-Dependent: NARCIS a National Research Classification in Isolation" In *28th ASIS SIG/CR Classification Research Workshop.* Advances in Classification Research Online 28, 7-10. https://journals.lib.washington.edu/index.php/acro/article/view/15393/12829

Smiraglia, Richard P. and Rick Szostak. 2018. "Converting UDC to BCC: Comparative Approaches to Interdisciplinarity." *Challenges and Opportunities for Knowledge Organization in the Digital Age: Proceedings of the Fifteenth International ISKO Conference, 9-11 July 2018, Porto, Portugal*, ed. Fernanda Ribeiro and Maria Elisa Cerveira. Advances in Knowledge Organization 16. Würzburg: Ergon, 530-8.

Szostak, Rick and Richard P. Smiraglia. 2017. "Comparative Approaches to Interdisciplinary KOSs: Use Cases of Converting UDC to BCC." In *Proceedings from North American Symposium on Knowledge Organization*, *Vol. 6. University of Illinois at Urbana-Champaign*, 202-15. https://journals.lib.washington.edu/index.php/nasko/article/view/15240/12698