Avoiding Research Tribal Wars Using Taxonomies

Steffen Wendzel, Worms University of Applied Sciences and FernUniversität in Hagen

Luca Caviglione, National Research Council of Italy

Wojciech Mazurczyk, Warsaw University of Technology

Every scientific domain benefits from a unified understanding and categorization of terms. This article highlights lessons learned from several years of taxonomy and terminology research in a cybersecurity domain.

A role of scientists that became increasingly important during recent years is one of public expounders. Phenomena like the COVID-19 pandemic, critical infrastructure security, and global warming are expected to be explained so that everybody can understand them. Moreover, such outreach activities are increasingly demanded by governments, funding bodies, and the public.

Another role of a scientific exchange is internal communication, where scientists discuss ideas, concepts, and results with their peers. However, even in science, there is a need for accessible communication, especially when it comes to interdisciplinary science, which can clash against cultural and language barriers as well as different technical backgrounds. This is, for instance, the case in human–computer interaction or machine learning-based cancer treatments where physicists work jointly with artificial intelligence experts. Similar cases include domain-overlapping areas, such as network security and systems security. In such cases, different subdomains use their own terms and taxonomies. With new concepts and ideas being introduced and the natural drive toward interdisciplinary research, novel terms emerge but overlap with existing ones. For instance, traditional Internet terms juxtapose with those of the Internet of Things (IoT) panorama, which are also used in Industrial IoT, Industry 4.0, or smart environments.

New terms should be created under consideration of their origin's meaning, which is often rooted in Latin or Greek.[1] When different domains develop their own

Digital Object Identifier 10.1109/MC.2022.3218175
Date of current version: 9 January 2023

understanding of terms, there is a risk that identical concepts (that is, scientific reinventions, i.e., ideas being rediscovered after their essence has already been introduced in the past) will use different terms in the long run, which can impinge on the understanding between these domains.[2] Such reinventions become more likely the more search terms are necessary to find related academic publications, and preventing them is becoming increasingly challenging as small research fields emerge, resulting in novel sub-sub disciplines, which then continue to grow autonomously. Donald Knuth already mentioned it in 2001.[3]

> "In my own field, for example, it once was possible for a grad student to learn just about everything there was to know about computer science. [...] Nowadays the subject is so enormous, nobody can hope to cover more than a tiny portion of it."

Now, 21 years later, with the advent of many more fields in information and communications technology [for example, augmented reality; software-defined networking (SDN); content delivery networks (CDNs); IoT, etc.], the situation is even worse, leading to very narrow specializations of research groups and practitioners. Therefore, we are experiencing an avalanche of separated research "islands" where different "tribes" speak diverse "languages." Researchers can even "fight" with each other instead of cooperating or trying to build upon a common understanding of some concepts or solving problems together. The "bigger picture" can often be missed or may be hard to comprehend, and differences in terminology make it even more confusing, thwarting the development of interdisciplinary research.

TAXONOMIES TO THE RESCUE

Minimizing scientific reinventions can be achieved via taxonomies, that is, ordered classification systems where a hierarchy can be used to describe when a category is derived from another. Despite sometimes being perceived as boring, they should be considered a major tool, and biology demonstrated that they can gain a key role in a scientific discipline. Before Swedish botanist Carl Linnaeus introduced his well-known taxonomy for biology,[4] several other taxonomies existed. The reason why his taxonomy became dominant mainly lies in its clarity and simplicity. Linnaeus provided a binomial nomenclature where species are named using only two components: the generic name and the specific name, for example, *Canis lupus* (gray wolf). Even if further expanded over the years, such a taxonomy remained applicable and attractive to its users.

If a taxonomy can handle the integration of novel ideas without restructuring, this should be considered a quality indicator.

LET'S NOT REINVENT THE WHEEL

We present some of the lessons learned while developing a taxonomy for the information hiding area.[5] The process is still not finished, but we believe some of our experiences can be useful for anyone with similar intents.

Creating and revising a taxonomy are usually nontrivial processes, especially when a research domain has emerged over decades. The more objects must be classified, the more challenging the generation of a suitable taxonomy is. In fact, taxonomies should be exhaustive and mutually exclusive[6]; that is, every object should be classifiable and should belong only to exactly one class.

New taxonomies should not necessarily reinvent a categorization from scratch but improve existing ones, driving an evolution rather than a revolution. This also lowers the chance for scientific reinventions as it prevents the introduction of too many novel terms and components. Of course, new additions can be introduced but only when necessary. If a taxonomy can handle the integration of novel ideas without restructuring, this should be considered a quality indicator. In other words, revisions should prioritize the clarity and comprehensiveness of the contained categories.

Another point to consider is that during the process of developing a coherent taxonomy spanning over several research subareas, experts from the relevant fields should be invited. Clearly, the more fields one wants to cover in the taxonomy, the more experts are needed. Managing such a team is a time-consuming effort as experts typically come from different backgrounds or use specific terminology, leading to members perceiving the same terms in a somewhat different way. The most suitable venues for presenting the outcome of works dealing with taxonomies can be journals and specific conferences, for instance, those considering systematization-of-knowledge submissions.

However, publishing a new taxonomy is just the first step to seek acceptance from the community. This usually requires quite a long time, effort, and a lot of basic work to let the community notice, understand, and appreciate the benefits. Efforts also include reaching out to potentially interested parties via social media; websites; invited talks; teaching activities like summer schools; etc.

STRIKING THE CORRECT BALANCE BETWEEN GENERIC AND SPECIALIZED TAXONOMIES

When taxonomies emerge over the years, they can drift further from related taxonomies' terms and categories. For this reason, it is helpful to merge related taxonomies from time to time. However, fusing domain-specific with domain-overlapping concepts into some major taxonomies might not always be feasible, especially for highly generic taxonomies. Mentionable examples of well-defined high-level taxonomies are IEEE's *2020 Taxonomy*[7] as well as ACM's *Computing Classification System*,[8] which suit the categorization of academic publications in computer science but not the detail-based categorization of very specialized subdomains (for this reason, publications can contain keywords).

In such a case, compatibility between more generic and more specialized taxonomies can still be sought by unifying the terminology whenever possible; for example, one could aim to unify the overlapping terms of SDN, CDN, or 5G taxonomy with some dominant taxonomy in computer networking. Classical pattern languages, such as the Pattern Language Markup Language (PLML),[9] can serve here as they allow one to introduce aliases (old terms must not be discarded and can still be mentioned as attributes, but both terms are described by the same pattern).

TAXONOMIES SHOULD KEEP SEMANTIC TECHNOLOGY IN MIND

Scientists often focus solely on categorizing and defining the links between terms. However, taxonomies may not be sufficient to keep scientific reinventions at the bare minimum. In fact, major challenges concern the creation, navigation, and sharing of knowledge, which account for additional tools. In this vein, the experience gained with the semantic web can be inspiring. Structured descriptions of data as provided by the Schema.org project and knowledge graphs with related data (such as Wikidata.org) or services such as the Google Knowledge Graph[10] should be considered as valuable resources for the definition of new taxonomies. Ultimately, taxonomies and semantic technologies should be developed jointly so as to limit the chance for reinventions originating from inconsistencies. Their elements should provide links and ease the discovery process for scientists investigating new or related areas with which they are not fully familiar yet. For instance, a network security taxonomy based on PLML (or even a Schema.org-like structure) might provide links to a knowledge graph that implements additional semantic detail of an object's relation and attributes.

Not surprisingly, taxonomies have been an important building block for the semantic Web and the definitions of the structures of large-scale datasets. In fact, "taxonomization" and the creation of ontologies are prime techniques to organize and navigate through the knowledge or a corpus of information. As happens for a search engine (see, for example, the use of the Google Knowledge Graph to enrich the search experience), taxonomies can be used to reveal new connections and to enlighten research areas apparently disjointed. In the case of scientific advancements or the definition of novel fields, they should not be considered only as a mere tool to classify ideas or concepts in an organized manner. Indeed, a taxonomy can facilitate the transfer of information or concepts but can also be a guideline to reveal links and overlaps that contradict common sense or are highly counterintuitive.

EXAMPLE: NETWORK STEGANOGRAPHY

In 2015, we published the first pattern-based taxonomy for network steganography,[11] which covered more than 100 hiding methods published since the 1980s. The initial taxonomy received an update in a textbook, followed by several extensions proposed during workshops, conferences, and in sections of journal articles. In the summer of 2021, ideas for the previously cited major revision of the taxonomy were proposed by a large consortium of experts from additional subdomains. This enabled us to maximize feedback before preparing a comprehensive overview of the steganography field.

In our latest taxonomy revision,[5] we extended our network steganography taxonomy to the whole domain of steganography, including digital media steganography; cyberphysical system/IoT steganography; file system steganography; and others. The coverage of this large research domain was not the original focus of this work and thus required several adjustments. However, several core aspects (parts of the terminology and structure) were kept, whereas others required more generalization. The revision of the taxonomy can now serve as a tool for keeping the terminology of the larger community consistent. In the future, the semantics of the research work is planned to become linked with the taxonomy using an ontology approach. However, also, for such advanced steps, a taxonomy can provide a solid foundation that streamlines succeeding steps.

Taxonomies can serve as an important tool for science-internal communication and can help to limit scientific reinventions. For this purpose, they can benefit from an evolutionary and structured development; should be kept accessible; and should not neglect taxonomies or terms of adjacent domains. As communities tend to push their own understanding of concepts, unified taxonomies can even help prevent "tribal wars" on the "correct" understanding of terms.

REFERENCES

1. P. Reviriego, E. Merino-Gòmez, and F. Lombardi, "Latin and Greek in computing: Ancient words in a new world," *Computer*, vol. 55, no. 6, pp. 66–72, Jun. 2022, doi: 10.1109/MC.2022.3149447.
2. D. K. Simonton, *Creativity in Science: Chance, Logic, Genius, and Zeitgeist*.

Cambridge, U.K.: Cambridge Univ. Press, 2004.
3. D. E. Knuth, *Things a Computer Scientist Rarely Talks About*. Stanford, CA, USA: CSLI, 2001.
4. C. Linnaeus. "Systema naturæ." Britannica. Accessed: Jul. 12, 2022. [Online]. Available: https://www.britannica.com/topic/Systema-Naturae
5. S. Wendzel et al., "A generic taxonomy for steganography methods," 2022, *TechRxiv PrePrint,* doi: 10.36227/techrxiv.20215373.
6. K. D. Bailey, *Typologies and Taxonomies: An Introduction to Classification Techniques*. Newbury Park, CA, USA: Sage, 1994.
7. "2022 IEEE Taxonomy," IEEE, Piscataway, NJ, USA, 2022. [Online]. Available: https://www.ieee.org/content/dam/ieee-org/ieee/web/org/pubs/ieee-taxonomy.pdf
8. "ACM computing classification system," Association for Computing Machinery, New York, NY, USA, 2012. [Online]. Available: https://dl.acm.org/ccs
9. S. Fincher, "PLML: Pattern Language Markup Language," CHI 2003 summary document, 2004. [Online]. Available: https://www.cs.kent.ac.uk/people/staff/saf/patterns/plml.html
10. "Google knowledge graph search API." Google. Accessed: Jul. 12, 2022. [Online]. Available: https://developers.google.com/knowledge-graph/
11. S. Wendzel, S. Zander, B. Fechner, and C. Herdin, "Pattern-based survey and categorization of network covert channel techniques," *ACM Comput. Surv. (CSUR)*, vol. 47, no. 3, pp. 1–26, 2015, doi: 10.1145/2684195.

STEFFEN WENDZEL is a professor and the scientific director at the Center for Technology & Transfer, Worms University of Applied Sciences, D-67549 Worms, Germany, and a lecturer at FernUniversität in Hagen, D-58097 Hagen, Germany. Contact him at steffen@wendzel.de.

LUCA CAVIGLIONE is a senior research scientist at the Institute for Applied Mathematics and Information Technologies of the National Research Council of Italy, I-16149 Genoa, Italy. Contact him at luca.caviglione@ge.imati.cnr.it.

WOJCIECH MAZURCZYK is a professor at the Warsaw University of Technology, 00-665 Warszawa, Poland. Contact him at wojciech.mazurczyk@pw.edu.pl

Digital Object Identifier 10.1109/MC.2022.3227705