Introducing an Evaluation Method for Taxonomies

Angelika Kaplan, Thomas Kühn, Sebastian Hahner, Niko Benkler
Jan Keim, Dominik Fuchß, Sophie Corallo and Robert Heinrich
firstname.lastname@kit.edu
niko.benkler@alumni.kit.edu
Karlsruhe Institute of Technology
Karlsruhe, Germany

ABSTRACT

Taxonomies are important for the development of a research field, as they play a major role in structuring a complex body of knowledge and help to classify processes, approaches, and solutions. While there is an increasing interest in taxonomies in the software engineering (SE) research field, we observe that SE taxonomies are rarely evaluated. To rise awareness and provide operational guidance on how to evaluate a taxonomy, this proposal-for-solution paper presents a three step evaluation method for taxonomies evaluating its structure, applicability, and purpose. This enables SE researchers to evaluate and improve the quality of their taxonomies and supports reviewers to assess the evaluation strategy.

KEYWORDS

taxonomies, evaluation, meta-research in software engineering

ACM Reference Format:
Angelika Kaplan, Thomas Kühn, Sebastian Hahner, Niko Benkler and Jan Keim, Dominik Fuchß, Sophie Corallo and Robert Heinrich. 2022. Introducing an Evaluation Method for Taxonomies. In *Evaluation and Assessment in Software Engineering (EASE 2022), June 13–15, 2021, Göteborg, Sweden.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Taxonomies[1] make a significant contribution to the organization and collection of knowledge in science and practice. They can serve to (1) classify objects of a research field, (2) provide a common terminology, (3) enable a better understanding of the interrelationships between the classified objects, (4) identify gaps, and (5) support decision making processes [22]. Although taxonomies are typically described as hierarchies of classes, taxonomies can have a wide variety of representations, e.g., a *hierarchy* with (or without) mutually exclusive classes, a *tree*, a *paradigm*, *facets*, a *ring*, or a *(knowledge) graph* [2, 18, 22]. Most taxonomies in Software Engineering (SE) are rather new and have the practical purpose of classifying among others processes, approaches, and solutions. More importantly, they help to cope with the increasingly huge body of knowledge in SE and foster understanding of complex SE domains [22]. Besides all that, taxonomies can enable communication among researchers and practitioners [23]. According to Usman et al. [22], there is an increasing interest in SE taxonomies, but they found a disturbing lack of evaluations of taxonomies, i.e., most only illustrated the taxonomy's utility (45.76%) or performed no validation at all (33.58%). However, evaluating taxonomies is no easy task, as quality criteria and comparable quality properties for taxonomies are hard to define and even harder to determine either quantitatively or qualitatively. Moreover, the evaluation should follow a systematic plan to structure its process accordingly. Furthermore, although current approaches (cf. Sect. 2) provide generic workflows, research methods, or techniques in a compendium, they do not give sufficient operational and practical guidance. Notably though, they already argued to consider the structural properties and utility as well as the purpose as all these aspects play a critical role for SE researchers to understand, apply, and compare taxonomies. To remedy this and to provide practical guidance, we introduce a method to systematically evaluate taxonomies in SE equally considering the suitability of its structure, its applicability, and its purpose. While taxonomies were typically evaluated through illustration or argumentation, we insist that a taxonomy should be evaluated by applying quantitative and qualitative metrics. Thus, we followed the Goal-Question-Metric (GQM) approach [4] as underlying structure to derive our goal-oriented evaluation method that maps the three evaluation goals to nine distinct quality criteria of taxonomies, which are measured by corresponding quantitative and qualitative metrics. In this paper, we address the following **research question**: *How to evaluate taxonomies in SE research and how to guide researchers through such an evaluation?* Our *contributions* are threefold: (1) We propose an initial method for evaluating taxonomies (in SE) with three successive steps that follows the GQM approach; (2) we provide operational guidance wrt. method design for a comprehensive taxonomy evaluation for both researchers and reviewers; and (3) we illustrate the method's application.

[1] Taxonomy—gr. *taxis* meaning order, arrangement; *nomos* meaning law or science.

EASE 2022, June 13–15, 2021, Göteborg, Sweden

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 STATE OF THE ART

Before introducing our evaluation method, we first discuss contemporary approaches for creating and evaluating taxonomies.

Approaches in Software Engineering Ralph [18] analyzes process theories and taxonomies in SE research to provide guidelines for their creation and evaluation comprising several steps including recommended research methods and guiding questions each, as depicted in Fig. 1. Usman et al. [22] defines four phases in the revised method for taxonomy development (extending Oré et al. [15]), as illustrated in Fig. 2. Specific activities are conducted in each

Generating Taxonomies	Evaluating Taxonomies
Step 1: Choose a Strategy	Step 1: Choose a Theory
Step 2: Site Selection	Step 2: Choose One or More Rival Theories
Step 3: Data Collection	Step 3: Analyze Thruth Claims
Step 4: Data Analysis	Step 4: Create Templates
Step 5: Conceptional Evaluation	Step 5: Select a Research Method
Step 6: Writing Up	Step 6: Collect and Analyze Data
Step 7: Peer Review	Step 7: Writing Up
	Step 8: Reviewing

Figure 1: Guidelines for taxonomies proposed by Ralph [18]

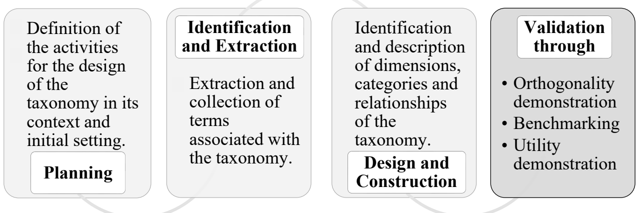

Figure 2: Revised taxonomy development method [22]

Table 1: Taxonomy Evaluation Framework, adapted from [20]

WHO? Subject of evaluation	Involvement	Subject has been involved in taxonomy building			Subject has not been involved in taxonomy building		
	Background	Academics			Practitioners		
	Experience	Domain			Method		
WHAT? Object of evaluation	Type	Real-world			About real-world		
	Involvement in taxonomy building	Object has been used in taxonomy building		Object has been used in taxonomy building and evolved since then		Object has not been used in taxonomy building	
	Coverage	Exhaustive		Selektive		Representative	
HOW? Method of evaluation	Approach	Quantitative			Qualitative		
	Method	Logical argument	Focus group	Expert interview	Survey	Delphi technique	Sorting
		Log diary	Case study	Illustrative scenario	Action research	Experiment	…

phase to create and evaluate a taxonomy in SE during development. However, the validation phase is restricted to orthogonality demonstration, benchmarking (i.e., comparison to similar taxonomies), and utility demonstration (i.e., demonstration by classifying subject matter examples). Still, for this phase, no corresponding metrics or guidelines were provided. In conclusion, both approaches provide hints regarding the process and workflow for generating and evaluating taxonomies, but concrete recommendations for the method design itself (e.g., definition of evaluation questions and criteria as well as corresponding metrics) are missing. They present template workflows and research method guidelines in a compendium style.

Approaches in Information Systems and Information Science Similar to Ralph [18], Szopinski et al. [20] regard taxonomies as a tool to analyze and understand a domain. In the domain of information systems, they also observe a lack of guidance for researchers on how to rigorously evaluate taxonomies. In turn, they present a framework, shown in Tab. 1, focusing on three main questions: how to evaluate, what to evaluate (object under study), and who evaluates (subject of evaluation). In a systematic literature review, Szopinski et al. [21] identify 54 papers that report on taxonomy

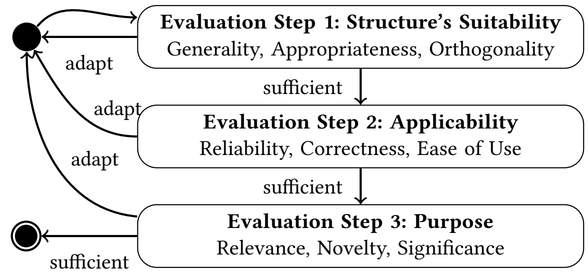

Figure 3: Overview of the process for evaluating taxonomies

Table 2: GQM-plan of our Method for Evaluating Taxonomies

Goal (Steps)	Question (Quality)	Metric
Suitable Structure	Generality Appropriateness Orthogonality	Laconicity, Lucidity Completeness, Soundness Orthogonality Matrix
Applicability	Reliability Correctness Ease of Use	Inter-Annotator Agreement Precision, Recall, F_1-Score Usability Score
Purpose	Relevance Novelty Significance	Fraction of Relevant Classes Innovation, Adaptation Classification Delta

evaluation criteria, i.e., quality criteria of a taxonomy, collecting 43 different evaluation criteria. Two evaluation criteria stand out as most frequently used: *usefulness* and *applicability*. Likewise, in Information Science, Bedford [2] identifies two evaluation issues with regard to classification schemes: evaluation of the classification scheme itself and evaluation of how well the scheme supports classification decisions. Both require their own framework and context for evaluation. Here, Ranganathan's principles of classification [19] are employed as framework for quality attributes, i.e., exclusiveness, uniqueness, relevance, ascertainability, consistency, affinity, decreasing extension, context, currency differentiation, exhaustiveness. In our approach, we aim to cover all of them.

Consequently, while each of the aforementioned research fields consider taxonomy evaluation criteria, there is no mapping to evaluation goals or metrics for data analysis. To remedy this, we explicitly map evaluation goals, to quality criteria representing evaluation questions or hypotheses and further to corresponding metrics.

3 TAXONOMY EVALUATION METHOD

In this section, we present our taxonomy evaluation method. Since a multitude of quality criteria have to be considered when evaluating taxonomies, our process distributes them over three steps in our evaluation process, as highlighted in Fig. 3. Our method applies the GQM-plan, shown in Tab. 2, to further structure and guide the evaluation process. We propose that in each of the three steps a specific *goal* is addressed, while a taxonomy's quality criteria correspond to the *questions* we ask. Finally, we align *metrics* to each quality criterion. For simplicity and without loss of generality, we consider a taxonomy to be a hierarchy of *categories* as nodes and *classes* as its leaves. As a running example, we consider a hierarchical taxonomy for mugs, sketched in Fig. 4 that classifies *mugs* wrt. their *size*, *material*, and intended *beverage*.

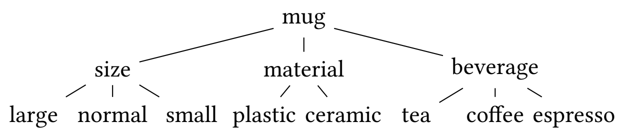

Figure 4: Tiny taxonomy for classifying different mugs

First, the taxonomy's *structure* is evaluated (cf. Sect. 3.1). As baseline, we check whether the taxonomy is suitable to classify objects under study, i.e., we evaluate its generality, appropriateness and orthogonality. If the structure is insufficient, the taxonomy has to be *adapted*, e.g., by defining more sound classes or removing unnecessary ones. Second, the taxonomy's *applicability* is evaluated (cf. Sect. 3.2). Here, we evaluate whether the taxonomy is usable and yields consistent results when employed by different users. We propose to carry out user studies to evaluate its reliability, correctness, and ease of use. Although user studies are able to identify structural problems of taxonomies, we included them as second step due to the required effort. If the applicability is insufficent, the taxonomy cannot be reliably applied and has to be improved based on the identified problems and user feedback. Third, the taxonomy's *purpose* is evaluated (cf. Sect. 3.3). In contrast to previous steps, evaluating the purpose of a taxonomy entails weighing its relevance wrt. preceding taxonomies. In detail, we evaluate its (internal) relevance, novelty, and significance, whereas the latter two only apply for preceding taxonomies with the same or a closely related purpose. If no such preceding taxonomies exist, one has to argue why existing taxonomies do not fit. In case the *purpose* is insufficient, the taxonomy must be further extended or an existing taxonomy should be employed instead. Please note that after any adaptation the taxonomy must be completely re-evaluated.

3.1 Step 1: Evaluating the Structure's Suitability

First, the suitability of the taxonomy's *structure* is evaluated. Since a taxonomy should permit the classification of objects under study, it must exhibit three structural quality criteria: *generality*, *appropriateness*, and *orthogonality*. To evaluate and quantify the *generality* and *appropriateness* of a taxonomy, we employ the four generalized metrics — *laconicity*, *lucidity*, *completeness*, and *soundness* — introduced by Ananieva et al. [1] (based on [7]). Although these metrics were developed to evaluate a conceptual model wrt. tools, we argue that they are also applicable to evaluate the generality and appropriateness of a taxonomy, since it abstracts from a set of objects under study with a dedicated purpose. In Evaluation Step 1, we only consider classes of the taxonomy that refer to terms of an object under study but not its categories. In case of the mug taxonomy, we only consider the leafs, e.g., large, normal, …, espresso.

Generality *Laconicity* and *lucidity* measure the generality of a taxonomy, i.e., whether it is both general and specific enough.

***Definition* 1** (Metrics for Generality)**.** *Let C be a taxonomy and $c \in C$ its classes, $\mathcal{R}$ a finite set of objects under study, $R \in \mathcal{R}$ an object under study with relevant terms $r \in R$, and $m_R^C \subseteq C \times R$ a relation between classes $c \in C$ and a relevant term $r \in R$. Then a term $r \in R$ is* laconic *wrt. a taxonomy C, if there is at most one c with $(c, r) \in m_R^C$. The corresponding function laconic(C, R, r) yields 1 if r is* laconic *and 0 otherwise. In turn, the* laconicity *metric is defined as:*

$$laconicity(C, \mathcal{R}) = \frac{\sum_{R \in \mathcal{R}} \sum_{r \in R} laconic(C, R, r)}{\sum_{R \in \mathcal{R}} |R|} \qquad \in [0, 1]$$

A class $c \in C$ is lucid*, if there is at most one $r \in R$ with $(c, r) \in m_R^C$. Conversely, the function lucid(C, R, c) yields 1 if c is* lucid *and otherwise 0. The* lucidity *metric is defined as:*

$$lucidity(C, \mathcal{R}) = \frac{\sum_{c \in C} \left(\min_{R \in \mathcal{R}} \ lucid(C, R, c) \right)}{|C|} \qquad \in [0, 1]$$

Laconicity determines the fraction of laconic terms among all objects under study, whereas a higher value is better. When classifying a portable cup with the mug taxonomy, the term *suitable for hot drinks* is **not laconic** as it is referenced by either espresso, coffee or tea. If two other terms where laconic, the resulting laconicity would be $2/3 \approx 0,66$ (only considering the portable mug). Low laconicity indicates that a class of the taxonomy may be too fine-grained, i.e., there are redundant classes in the taxonomy that should be merged. Similarly, *lucidity* determines the fraction of lucid classes, which, in turn, should approach one. In case of the mug taxonomy, the class *plastic* might **not be lucid** wrt. the distinct terms *made from PET* and *contains Polystyrene*. If all other classes were lucid, the lucidity of the mug taxonomy would be $7/8 \approx 0,88$. Low lucidity entails that classes of the taxonomy are too coarse-grained, meaning that there are unspecific classes in the taxonomy that should be split up. In sum, a taxonomy has a suitable generality, if both laconicity and lucidity are sufficiently high.

Appropriateness Similarly, *completeness* and *soundness* assess the taxonomy's *appropriateness*, i.e., whether it fully and correctly covers all relevant terms of the objects under study.

***Definition* 2** (Metrics for Appropriateness)**.** *Under the preconditions of Def. 1, a term $r \in R$ is* complete*, if there is at least one $c \in C$ with $(c, r) \in m_R^C$. The corresponding function complete(C, R, r) yields 1 if r is* complete *and 0 otherwise. The* completeness *metric is defined as:*

$$completeness(C, \mathcal{R}) = \frac{\sum_{R \in \mathcal{R}} \sum_{r \in R} complete(C, R, r)}{\sum_{R \in \mathcal{R}} |R|} \qquad \in [0, 1]$$

Likewise, a class $c \in C$ is sound*, if there is at least one $r \in R$ with $(c, r) \in m_R^C$. The function sound(C, R, c) yields 1 if c is* sound *and otherwise 0. The* soundness *metric is defined as:*

$$soundness(C, \mathcal{R}) = \frac{\sum_{c \in C} \left(\max_{R \in \mathcal{R}} \ sound(C, R, c) \right)}{|C|} \qquad \in [0, 1]$$

Completeness denotes the fraction of complete terms over all objects under study. Regarding the portable cup, the term *closable lid* is *not complete*, as it is not covered by a class in the mug taxonomy. If all three other terms were complete, i.e., covered by at least one class, the completeness would be $3/4 = 0,75$ (for the portable cup). Low completeness reveals that the taxonomy lacks classes that should be added to cover all relevant terms. By contrast, *soundness* represents the fraction of sound classes in the taxonomy. For example, the class *ceramic* is **not sound** when only considering portable cups, resulting in a soundness of $7/8 \approx 0,88$ if all other classes were sound. Low soundness indicates that the taxonomy may include unnecessary classes that can be removed or that the objects under study lack diversity. In conclusion, a taxonomy is appropriate, if it has both sufficiently high completeness and soundness.

Table 3: Orthogonality Matrix for the Tiny Mugs Taxonomy (1 indicates a dependence and 0 none).

	small	medium	large	plastic	ceramic	espresso	coffee	tea	$\sum$
small	-	0	0	0	0	**1**	0	0	**1**
medium	0	-	0	0	0	0	0	0	**0**
large	0	0	-	0	0	0	0	**1**	**1**
plastic	0	0	0	-	0	0	0	0	**0**
ceramic	0	0	0	0	-	0	0	0	**0**
espresso	**1**	0	0	0	0	-	0	0	**1**
coffee	0	0	0	0	0	0	-	0	**0**
tea	0	**1**	**1**	0	0	0	0	-	**2**
$\sum$	**1**	**1**	**1**	**0**	**0**	**1**	**0**	**1**	**5**

Orthogonality For taxonomies, Bedford [2] argues that *"No two categories should overlap or should have exactly the same scope and boundaries"* and, as such, requires *orthogonality* among a taxonomy's classes. To evaluate orthogonality, we can employ a self-referencing *orthogonality matrix*, where the classes of a taxonomy denotes the columns and rows of the matrix (cf. [16]). Then, individual cells are filled with either zeroes, if two classes are independent, or a positive number, if the class in its row implies the class in its column. Tab. 3 illustrates a possible orthogonality matrix for the mug taxonomy, where an espresso implies a small cup and vice versa. Tea implies a medium or a large mug, but only a large mug implies tea. These dependencies can be either perceived dependencies or computed functional dependencies (cf. [8]) between all classes wrt. the classification of all objects under study. Consequently, fewer dependencies indicate an improved orthogonality.

3.2 Step 2: Evaluating Applicability

Second, the taxonomy's applicability is evaluated. A taxonomy needs to be understandable and usable to be applicable [2]. This can be shown by means of user studies. In case user studies are not feasible, researchers can demonstrate the applicability by themselves using case studies (cf. Sect. 4). In general, a taxonomy's applicability is determined by its *reliability* (i.e., consistency of results), *correctness* of results, and *ease of use*.

Reliability To evaluate the taxonomy's reliability, one has to show that different users come to the same or at least very similar results for a classification task. For this, a user study needs to be conducted, where users apply the taxonomy on the same objects under study. The resulting classifications are then compared between users.

Metrics for Reliability. For this comparison, different metrics can determine the inter-annotator agreement (inter-rater reliability). In case of the mug taxonomy, one would task separate groups to classify each mug of a specific catalog. Afterwards, the inter-annotator agreement is determined by the overall overlap of the individual classifications for each mug in the catalog. Commonly used metrics are Cohen's κ [5], Fleiss' κ [6], and Krippendorff's α [11]. Each of them has specific benefits and drawbacks. Cohen's κ permits calculating the agreement between pairs of annotators, Fleiss' κ and Krippendorff's α for any number of annotators. Cohen's and Fleiss' κ work best with large sample sizes. In general, Krippendorff's α is a good choice as it is very flexible and can deal with things such as incomplete data, varying sample sizes and various categories. The results of these metrics need to be interpreted. An acceptable level of agreement depends on the domain and application. According to Krippendorff [10], it is customary to require a value above 0.80. Similarly, Landis and Koch [12] define that a value between 0.41 and 0.60 shows a moderate agreement, a value between 0.61 and 0.80 shows a substantial agreement, and a value above 0.80 shows almost perfect agreement.

Correctness Although users might produce reliable results, they can still differ from the intended results, e.g., if class definitions are not clearly defined or are ambiguous. Thus, comparing taxonomies from user studies with a gold standard is necessary.

Metrics for Correctness. Metrics like precision, recall, and F_1-score can help to show how correct a taxonomy was applied. In case of the mug taxonomy, several mugs could be classified by experts creating the gold standard, and then again by multiple different users. Succinctly, the users' precision, recall, and accuracy can be determined. Researchers can calculate most of these metrics for single classes to gain deeper insights, i.e., over- or underperformance of certain classes. To calculate an overall performance score of correctness, we argue that researchers should compute both overall and weighted averages. While the former averages all performance scores of each class, the latter is weighted with the number of occurrences. Both together will better indicate the taxonomy's correctness.

Ease of Use Besides a taxonomy's reliability and correctness, its *ease of use* can also be evaluated indicating whether users are able to understand and apply the taxonomy easily.

Metrics for Ease of Use. Researchers have various options to evaluate usability, e.g., asking users via questionnaires [9] or observing users during the classification task by utilizing the think-aloud technique [24]. Besides all that, we recommend to apply and adapt standards, such as the System Usability Score (SUS) [13]. For our mug taxonomy, we would simply ask the participants of the correctness study to complete a usability questionnaire afterwards.

3.3 Step 3: Evaluating Purpose

Finally, the taxonomy's semantics and relation to previous taxonomies are evaluated. To this end, the taxonomy's *relevance*, *novelty*, and *significance* are measured. In this step, both classes and categories of a taxonomy must be considered.

Relevance This evaluation considers whether each individual class and category provides value for the taxonomy's purpose. In general, a taxonomy should only cover classes that are relevant for the objects under study and the taxonomy's purpose. In particular, a taxonomy should not contain unnecessary or superficial classes for its designated purpose. For the mug taxonomy, a category *material density* might not be a valuable addition to help customers of a webshop to distinguish mugs.

Metrics for Relevance. Deciding whether or not a class or category is relevant depends on its individual semantics. Explicitly distinguishing between relevant and irrelevant classes (and categories) yields the *fraction of relevant classes and categories* as corresponding metric. The mug taxonomy, e.g., contains the classes *coffee*, *tea*, and *espresso*, which might not serve the purpose of distinguishing mugs, who typically only distinguish between their suitability for hot or cold beverages. Thus, the mug taxonomy has a fraction of relevant classes/categories of $9/12 = 0,75$. A low fraction indicates the existence of irrelevant classes or categories that should be removed.

Novelty While relevance only considers the currently evaluated taxonomy, its *novelty* is relative to previous taxonomies. It is an indicator for the *innovation* and *adaptation* of the evaluated taxonomy when compared to previous taxonomies with a similar purpose. While *innovation* depends on the newly introduced classes and categories, *adaptation* takes existing classes and categories into account, whose semantics have been adapted to fit the desired purpose.

***Definition* 3** (Metrics for Novelty)**.** *Let C be a taxonomy of classes and categories $c \in C$; $\mathcal{T}$ a finite set of previous taxonomies $T \in \mathcal{T}$ with classes and categories $d \in T$; and $\simeq \subseteq C \times T$ denote that a class/category $c \in C$ is* adapted from *a class/category $d \in T$ (written as $c \simeq d$) whereas $c \neq d$ holds. A class/category $c \in C$ is* new, *if $c \neq d$ and $c \not\simeq d$ for all $d \in T$. The corresponding function $new(C, T, c)$ yields 1 if c is* new *and 0 otherwise. Then,* innovation *is defined as:*

$$innovation(C, \mathcal{T}) = \frac{\sum_{c \in C} \min_{T \in \mathcal{T}} new(C, T, c)}{|C|} \qquad \in [0, 1]$$

Similarly, a class/category $c \in C$ is adapted, *if $c \simeq d$ for any $d \in T$. The corresponding function $adapted(C, T, c)$ yields 1 if c is* adapted *and 0 otherwise. Then,* adaptation *is defined as:*

$$adaptation(C, \mathcal{T}) = \frac{\sum_{c \in C} \max_{T \in \mathcal{T}} adapted(C, T, c)}{|C|} \qquad \in [0, 1]$$

Note that $0 \leq innovation(C, \mathcal{T}) + adaptation(C, \mathcal{T}) \leq 1$ holds for arbitrary taxonomies C and finite sets of taxonomies $\mathcal{T}$.

We consider a class or category to be *adapted from* a previously existing class or category, if a name, semantics or position was adapted. A class or category in the evaluated taxonomy is counted as *new*, if none of the previous taxonomies contain the same or an adapted class or category. In contrast, a class or category is counted as *adapted*, if there is at least one previous taxonomy containing an adapted class or category. For example, we would extend the mug taxonomy refining the class *plastic* into the category *plastic* containing two new classes *PET* and *PPE*. In comparison, this extended taxonomy would yield an *innovation* of $2/14 \approx 0,14$ and *adaptation* of $1/14 \approx 0,07$. The sum of *innovation* and *adaptation* indicates the overall novelty of the evaluated taxonomy. Please note, that depending on its purpose the lower innovation and/or lower adaptation might suffice to support taxonomy's significance, e.g., if the taxonomy's purpose is to combine different taxonomies it should still be considered sufficiently novel.

Significance. We consider a taxonomy to be more significant than others, if it enables a more detailed categorization of objects under study. We can only compare taxonomies with the same purpose by applying them on a common set of objects. In general, we compare the number of equivalence classes of the evaluated taxonomy with those of previous taxonomies.

***Definition* 4** (Metrics for Significance)**.** *Let C be a taxonomy of classes and categories $c \in C$; $\mathcal{T}$ a finite set of previous taxonomies $T \in \mathcal{T}$ with classes and categories $t \in T$; $\mathcal{R}$ be a finite none empty set of objects under study; and $\sim_T \subseteq \mathcal{R} \times \mathcal{R}$ denotes an equivalence relation for a taxonomy $T \in \mathcal{T} \cup \{C\}$, whereas $\sim_T$ denotes that a pair of objects is classified identically wrt. taxonomy T. Then, the* classification delta *over $\mathcal{R}$ is defined as:*

$$classification\ delta(C, \mathcal{T}, \mathcal{R}) = \frac{|\sim_C| - (\max_{T \in \mathcal{T}} |\sim_T|)}{|\mathcal{R}|} \qquad \in [-1, 1]$$

The *classification delta* determines the normalized difference between the number of equivalence classes between the evaluated taxonomy and the most detailed one. A positive result indicates a more detailed taxonomy, as it improves the distinction between objects under study. In contrast, a negative result suggests that a more detailed taxonomy exists that could be used instead. Besides that, if the delta is zero the taxonomy might still be sufficiently different, yet might be improved by including categories and classes of the most significant taxonomy. In case, the mug taxonomy would yield 5 equivalence classes for 5 PET and 5 PPE mugs. Then, the extended mug taxonomy (with PET and PPE as sub-classes of plastic) would yield 10 equivalence classes, resulting in a classification delta of $(10 - 5)/10 = 0.5$. Conversely, the extended mug taxonomy permits a more detailed categorization and is thus more significant.

4 ILLUSTRATIVE APPLICATION

In a separate publication an author proposed a novel taxonomy to classify and distinguish uncertainties in software architectures. (Reference will be ready for camera-ready version.) The purpose of this taxonomy is to enable software architects to structurally distinguish types of uncertainties and estimate their impact on software architectures. It is hierarchically structured with 10 categories below the root and 32 classes as leafs. In the publication, the taxonomy is applied on the architecture documentation of an open-source contact tracing app (CWA) to extract and classify several types of uncertainties. Besides that, the author applied the presented evaluation process for this taxonomy.

Regarding the **structure's suitability**, the author evaluated the *generality* and *appropriateness* wrt. the uncertainties found in the documentation, yielding a laconicity of 1.0 and a lucidity of 1.0 as well as a completeness of 0.97 and a soundness of 0.97. The taxonomy's *orthogonality* was not evaluated with an orthogonality matrix. Instead, the author analyzed the found and classified types of uncertainty and argued that the classes are independent wrt. their statement of impact. Consequently, the evidence for the taxonomy's orthogonality is unstructured and potentially incomplete. Next, the taxonomy's **applicability** was demonstrated. The author showed how the *uncertainty* taxonomy can be applied to extract uncertainties from the documentation and estimate their impact on the software architecture. However, as the author did not conduct a user study, it is impossible to make statements about the taxonomy's reliability, correctness, and ease of use. While the applicability was demonstrated, the author deemed it sufficient enough to finally evaluate the taxonomy's **purpose**. First, the *relevance* of each category and class was justified to help distinguish types of uncertainties and estimate their impact resulting in a fraction of relevant classes and categories of 1.0. Next, the *novelty* of the taxonomy was determined wrt. three existing taxonomies for uncertainty [3, 14, 17]. The overall innovation is 0.59 and the adaptation is 0.22, highlighting that most of the classes were adapted or newly created to serve the taxonomy's purpose. Last but not least, the taxonomy's *significance* was measured relative to the three aforementioned taxonomies. In fact, all three were applied to classify the

28 uncertainties described in the CWA case. Comparing the classification results, the best previous taxonomy, i.e, [17], produced 8 equivalence classes, whereas the author's taxonomy yielded 21. The *classification delta* is approx. 0.46, indicating a considerable improvement relative to the three previous taxonomies.

Although this illustrative application relied on argumentation instead of a user study to determine the taxonomy's applicability, this example showcases the different steps in practice. Moreover, it emphasizes the effect that omitted evaluation steps introduce threats to validity, e.g., generalizability or replicability.

5 DISCUSSION

Finally, we answer our research question, consider limitations of our evaluation method, and discuss threats to validity for this proposal.

RQ: How to evaluate taxonomies in SE research and how to guide researchers through such an evaluation? Instead of an unstructured evaluation, we follow a GQM-approach indicating a taxonomy's structure, applicability, and purpose as evaluation goals. These are mapped to three distinct quality attributes each. To evaluate each attribute, we provide and discuss corresponding metrics. The metrics provide evidence for the quality attributes, which, in turn, indicate the sufficiency of the taxonomy's structure, applicability, and purpose. While it is possible to provide arguments as evidence for some quality attributes, this introduces threats to validity to the evaluation. The described GQM-plan, process, and metrics provide the structure and means for the practical application of the proposed taxonomy evaluation method.

Limitations The completeness of quality attributes and metrics is a major concern. Conducting further case studies and literature reviews will help to complement missing aspects. Furthermore, template questions and hypotheses for each quality criterion were not considered, yet. Additionally, our GQM-based method does not provide guidance for a specific research method, instead we mainly define a dedicated goal for each step of the evaluation. For each goal, we declare relevant quality attributes and corresponding metrics for operational guidance. Granted, all these elements can also be embedded within a specific research method. Besides all that, as a proposal-for-solution, we can only provide an initial approach for evaluating taxonomies and an illustrative example of its applicability. However, we concede that more rigorous evaluations with several case studies are still needed.

Threats to Validity wrt. Method Design Although our method for evaluating taxonomies intend to be generalizable (*external validity*) in the SE research context, we assume that it is applicable to other research fields as well. Regarding *internal validity*, we ensured that the granularity of each step and metric is adequate and is independently focused on either a taxonomy's structure, applicability, or purpose. Moreover, each quality criterion and its corresponding metrics can be independently compiled. The only dependencies are enforced by consecutive order of the evaluation steps, such that, e.g., a weak evaluation of the applicability (Step 2), will also weaken and threaten the evaluation of the purpose (Step 3). Finally, *construct validity* was ensured by following the GQM-approach mapping goals to questions (in our case qualities) and further to dedicated metrics. These provide quantitative and qualitative measurements for a proper analysis of evaluation results.

6 CONCLUSION

In this paper, we proposed a GQM-based method for evaluating taxonomies in SE consisting of three consecutive steps focusing on the *structure*, *applicability*, and *purpose* of a taxonomy. For each step, we define quality attributes and corresponding metrics. We showcased the application of this method within SE research highlighting limitations when deviating from our method. Thus, both researchers and reviewers can benefit from the proposed evaluation method. In ongoing work, we plan to refine and consolidate the definition of our metrics and corresponding questions (or hypotheses) to the quality criteria and evaluate them in multiple case studies.

REFERENCES

[1] Sofia Ananieva et al. 2020. A Conceptual Model for Unifying Variability in Space and Time. In *Int. Conf. on Sys. & Softw. Product Lines*, Vol. A. ACM, Article 15, 12 pages.

[2] Denise Bedford. 2013. Evaluating classification schema and classification decisions. *Bulletin of the American Society for Information Science and Technology* 39, 2 (2013), 13–21.

[3] Tomas Bures et al. 2020. Capturing Dynamicity and Uncertainty in Security and Trust via Situational Patterns. In *Leveraging Applications of Formal Methods, Verification and Validation: Engineering Principles*. Springer, 295–310.

[4] Victor R. Basili Gianluigi Caldiera and H. Dieter Rombach. 1994. The goal question metric approach. *Encyclopedia of Softw. Eng.* (1994), 528–532.

[5] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46.

[6] Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin* 76, 5 (1971), 378.

[7] Giancarlo Guizzardi et al. 2005. An Ontology-Based Approach for Evaluating the Domain Appropriateness and Comprehensibility Appropriateness of Modeling Languages. In *Int. Conf. on Model-Driven Eng. Languages & Systems (MODELS)*. Springer.

[8] Jan L. Harrington. 1995. *Relational Database Design*. Prentice Hall Austria.

[9] Andrew Hodrien, TP Fernando, et al. 2021. A review of post-study and post-task subjective questionnaires to guide assessment of system usability. *J. of Usability Studies* 16, 3 (2021), 203–232.

[10] Klaus Krippendorff. 2004. Reliability in content analysis: Some common misconceptions and recommendations. *Human Comm. Research* 30, 3 (2004), 411–433.

[11] Klaus Krippendorff. 2018. *Content Analysis: An Introduction to its Methodology*. Sage Publishing.

[12] J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 1 (1977), 159–174.

[13] James R Lewis. 2018. The system usability scale: past, present, and future. *Int. J. on Human-Computer Interaction* 34, 7 (2018), 577–590.

[14] Sara Mahdavi-Hezavehi et al. 2017. A classification framework of uncertainty in architecture-based self-adaptive systems with multiple quality requirements. In *Managing Trade-Offs in Adaptable Software Architectures*. Elsevier, 45–77.

[15] Sussy Bayona Oré et al. 2014. Critical success factors taxonomy for software process deployment. *Softw. Qual. J.* 22, 1 (2014), 21–48.

[16] Karen J. Ostergaard and Joshua D. Summers. 2009. Development of a systematic classification and taxonomy of collaborative design activities. *J. of Eng. Design* 20, 1 (2009), 57–81.

[17] Diego Perez-Palacin and Raffaela Mirandola. 2014. Uncertainties in the Modeling of Self-Adaptive Systems: A Taxonomy and an Example of Availability Evaluation. In *5th Int. Conf. on Performance Eng. (ICPE)*. ACM, 3–14.

[18] Paul Ralph. 2019. Toward Methodological Guidelines for Process Theories and Taxonomies in Software Engineering. *IEEE Trans. on Softw. Eng.* 45, 7 (2019), 712–735.

[19] Shiyali Ramamrita Ranganathan. 1937. *Prolegomena to library classification*. Madras Library Association.

[20] Daniel Szopinski et al. 2019. Because Your Taxonomy is Worth IT: towards a Framework for Taxonomy Evaluation. In *27th European Conf. on Information Systems*. Association for Information Systems, 19.

[21] Daniel Szopinski et al. 2020. Criteria as a Prelude for Guiding Taxonomy Evaluation. In *53rd Hawaii Int. Conf. on Sys. Sciences*. ScholarSpace, 1–10.

[22] Muhammad Usman et al. 2017. Taxonomies in software engineering: A Systematic mapping study and a revised taxonomy development method. *Information and Software Technology* 85 (2017), 43–59.

[23] Sira Vegas et al. 2009. Maturing Software Engineering Knowledge through Classifications: A Case Study on Unit Testing Techniques. *IEEE Trans. on Softw. Eng.* 35, 4 (2009), 551–565.

[24] Lev S. Vygotsky. 1962. *Thought and Language*. MIT Press.