Testing the validity of Wikipedia categories for subject matter labelling of open-domain corpus data

Ahmad Aghaebrahimian*
Andy Stauder
Michael Ustaszewski*

Department of Translation Studies, University of Innsbruck, Austria

Abstract
The Wikipedia category system was designed to enable browsing and navigation of Wikipedia. It is also a useful resource for knowledge organisation and document indexing, especially using automatic approaches. However, it has received little attention as a resource for manual indexing. In this article, a hierarchical taxonomy of three-level depth is extracted from the Wikipedia category system. The resulting taxonomy is explored as a lightweight alternative to expert-created knowledge organisation systems (e.g. library classification systems) for the manual labelling of open-domain text corpora. Combining quantitative and qualitative data from a crowd-based text labelling study, the validity of the taxonomy is tested and the results quantified in terms of interrater agreement. While the usefulness of the Wikipedia category system for automatic document indexing is documented in the pertinent literature, our results suggest that at least the taxonomy we derived from it is not a valid instrument for manual subject matter labelling of open-domain text corpora.

Keywords
Corpus labelling; social tagging; taxonomy; Wikipedia

1. Introduction
Being one of the largest knowledge resources and crowd-based endeavours on the web to date, Wikipedia has been studied extensively in numerous disciplines, including information science, linguistics, computer science and natural language processing, to name but a few. Its size, diverse topics, multilinguality, free availability, collaborative nature, dynamic growth, openness and transparency make it an unprecedented social and technological phenomenon that yields rich data for a wide range of research questions. Wikipedia-based research mainly falls into two categories: first, research into the Wikipedia phenomenon itself, and, second, studies using Wikipedia as a data source for other research interests and applications [1,2].

A prolific line of inquiry from the second category deals with the use of Wikipedia for knowledge acquisition and knowledge organisation. More specifically, the community-driven Wikipedia category system (WCS) for subject matter indexing is being explored as an alternative to more traditional knowledge organisation systems created by dedicated expert groups, or expert-created systems for short. The WCS is conceived by the Wikipedia community as a polyhierarchy of navigational links for browsing Wikipedia pages [3]. While by 2012 the WCS had received relatively little attention [4], a recent bibliographic review [5] has identified a large body of research on the exploration and application of the WCS, highlighting its suitability for various problems in knowledge organisation. Large part of the reviewed works deal with automatic and manual indexing and categorisation. This article is in the vein of these works. Given its comprehensive coverage, the WCS can be used to index collections of data belonging to different types of media (see Section

*Equal contribution.

Corresponding author:
Michael Ustaszewski, Department of Translation Studies, University of Innsbruck, Herzog-Siegmund-Ufer 15, A-6020 Innsbruck, Austria.
Email: michael.ustaszewski@uibk.ac.at
3). While most studies and applications use knowledge organisation systems derived from WCS for domain-specific problems [5], this article aims to evaluate the suitability of a reduced version of the WCS for open-domain problems. Scholarly interest in the WCS may partly be attributed to the growing need to assign descriptive metadata to the vast amount of unstructured and/or unlabelled data in an increasingly data-centric world. While considerable research has been carried out on automatic approaches, evaluating the validity of WCS for manual indexing has as yet received little attention. This is somewhat surprising, since automatic indexing can only produce useful results if the underlying vocabulary organises concepts in a meaningful, intersubjective and manageable way, which are inherently features of intellectual metadata (a term employed by Greenberg et al. [6]). To address this gap, this article explores whether the WCS can be used as a resource to build an intuitive lightweight taxonomy for crowd-based manual document labelling by non-experts. Intuitive labelling tools are vital to collaborative corpus building initiatives, such as the open-access translation repository TransBank - A Meta-Corpus for Translation Research [7], in whose context this study was carried out.

2. Aims

The WCS is a crowd-based social tagging system that has grown organically rather than an expert-created system. Social tagging in general comes with some potential drawbacks, such as lack of precision, ambiguities and inconsistencies [8,9]. Conflicting assumptions about fundamental design choices among editors and disregard of theories from the knowledge organisation literature resulted in structural inconsistencies with regard to the hierarchical organisation of the WCS [10]. However, these phenomena, which can be subsumed under the category of meta noise, are offset by the aforementioned features of intellectual metadata, provided that the number of contributing users is large enough [11]: ‘[I]disyncratic tagging, or meta noise, tends to have a negligible impact as the number of users get large. In fact these design solutions are more robust with a larger amount of underlying data’ (p. 113).

Against this background, the goal of this study is to bridge the gap between two use cases:

1. Browsing, searching and navigating knowledge bases: for this purpose, indexing systems have (been) developed, which are mostly complex and comprehensive in the number of topics they cover.
2. Browsing, searching and navigating text corpora: for this purpose, mostly dedicated and coarse-grained label sets with only few categories have been developed. Highly detailed subject matter labelling is usually not a priority for these systems.

This article concerns itself with the evaluation of the potential transfer of an indexing system from Use Case 1 to Use Case 2: the usefulness of the WCS for subject matter labelling of open-domain corpora is to be evaluated with the help of a crowd-based interactive online study. The reason for selecting such an approach is the similarity of knowledge bases and text data: texts (almost) always contain knowledge, and knowledge is mostly stored in the form of text. Use Case 1, of course, contains more and other types of systems than the WCS, such as library classification systems (e.g. Universal Decimal Classification, Dewey Decimal Classification) or thesauri. However, we selected WCS for evaluation because it combines a number of potential advantages:

- Intersubjectivity: categories generated based on community effort and negotiation among editors;
- Intuitiveness of the hierarchical relations, as shown by previous research [12]: this may be due to the aforementioned intersubjectivity advantage;
- Ease of use for non-experts: created for the general public;
- Categories refer to subject matters and do not intermix these with genres and text types, as opposed to, for example, library classification systems.

In addition to these advantages, which make the WCS a promising candidate for a labelling tool transfer from Use Case 1 to Use Case 2, the complexity of the system was reduced prior to evaluation (see Section 4.1): for the sake of usability, we aim to strike a middle ground between extremely detailed, fine-grained systems (e.g. library classification systems) on one hand, and coarse-grained and therefore easy-to-use systems (e.g. subjects domains of British National Corpus3 or BNC) on the other hand.

Summing up these initial considerations, in this study, we conduct an extrinsic task-based evaluation of a hierarchical taxonomy built from a subset of the WCS. The study combines quantitative and qualitative data from an open-domain text corpus labelling task to test the validity of the derived taxonomy, operationalised through interrater agreement (IRA).
3. Related work

Numerous studies use the WCS for knowledge organisation or explore it as an object of study in its own right. The following is an overview of selected studies that share some goals with this study and/or have important implications for its design. According to our literature review, there is no other work that evaluates the validity of applying a taxonomy derived from WCS as a controlled vocabulary, that is, a structured arrangement of terms for indexing content and retrieving content through browsing and searching [13], for crowd-based subject matter labelling.

Kotlerman et al. [14] use the WCS to build a domain-specific taxonomy for the indexing of video content by combining automatic procedures with human post-editing. As a preprocessing step, cycles from the WCS graph are removed. In contrast, Ponzetto and Strube [15] build a domain-independent general-purpose taxonomy from WCS, which achieves similar quality to expert-created gold-standard knowledge bases. The extrinsic task-based evaluation uses the induced taxonomy to compute semantic similarity scores between pairs of words, whereas our study builds a taxonomy for the manual labelling of text documents.

Josi et al. [16] present an automatic method that uses WCS as a controlled vocabulary for indexing scientific images, based on the images’ captions and textual fragments referring to the images. They use almost the entire category graph rather than a reduced subset. The method is evaluated extrinsically on 100 ground-truth images manually labelled by non-experts. In our evaluation, the manual labelling is performed on text documents using a custom taxonomy built from a subset of the category graph.

TagTheWeb [17] is a general-purpose tool for the automatic categorisation of text documents from any domain. For a given document, the tool calculates the relative influence of each of the 19 top-level Wikipedia categories in the document. The tool is conceptually similar to our study in that it aims at open-domain problems and in that it reduces the complexity of the WCS graph by focusing on a limited number of categories and levels of the graph. However, by using top-level categories only in what they call text ‘finger prints’, the reduction is much more drastic than in our study.

Several studies analyse the structure and composition of the WCS graph rather than using it as a resource for indexing. Voß [18] describes the WCS as a collaboratively constructed thesaurus with hierarchical relations, which can be used to derive hierarchical tree structures similar to expert-created classification systems such as the Dewey Decimal Classification. Bairi et al. [19] diachronically analyse the distribution of Wikipedia articles, with a focus on subsumption relations and the presence of cycles in the WCS graph, concluding that up to a depth of three to five levels, the graph yields reasonably good subsumption relations.

Fernando et al. [12] compare the usefulness of six knowledge organisation systems, among them one taxonomy derived from the WCS [15], for the organisation of large cultural heritage collections. The crowd-based intrinsic evaluation measures the cohesion of the six systems, that is, how well similar concepts are grouped, and examines how understandable relationships between concepts are. With regard to concept relations, the WCS-based taxonomy was the best performing system, suggesting that the socially constructed WCS is of high quality and has easy-to-understand concept relations. The expert-created Library of Congress Subject Headings scored much worse, which suggests that its concept relations are harder to understand and identify for users. This finding is crucial for the purpose of our crowd-based study, because it indicates that library classification systems, which are usually applied by trained professionals, are less intuitively understandable than socially constructed knowledge organisation systems and may therefore be less suitable for non-expert indexing tasks. In terms of cohesion, the WCS-based taxonomy was the third best system and of similar quality to the Library of Congress Subject Headings, both of which were outperformed by an automatically generated taxonomy based on Wikipedia article links.

Boldi and Monti [20] have developed a method for the denoising and pruning of the WCS graph. The method is proposed as a preprocessing step in automated tasks that extract knowledge from WCS, such as ontology building. While their work aims to improve the quality of the WCS using a graph-theoretic approach, our study aims to reduce its complexity and to assess its validity in a task-based manual labelling study.

Comparisons of the WCS with expert-created classification systems have shown that WCS has become more similar to the latter, especially with regard to the top-level categories [4]. Thus, most of Wikipedia’s top-level categories have a clear counterpart in the Universal Decimal Classification [21]. Such findings contributed to the conclusion that the WCS is a viable alternative to closed-context, expert-generated knowledge organisation systems [5].

4. Materials and methods

4.1. Derivation of hierarchical taxonomy from Wikipedia categories

Hierarchical taxonomies offer flexible degrees of granularity in subject matter labelling, because they let annotators choose between specific labels from deeper levels of the taxonomy and less specific labels from higher levels, depending
on how sure they are about the subject of a given document. In case of uncertainty, more general but accurate rather than specific but incorrect labels should be used [13] (p. 174). For crowd-based corpora labelled by non-experts, this flexibility helps avoid labelling errors, because annotators are not forced to assign specific labels when in doubt. Furthermore, hierarchies are very well suited for navigation and browsing [10]. In combination with its size and coverage, its hierarchical structure makes the WCS a prime candidate for the labelling and navigation of open-ended crowd-based text collections. However, the WCS is not without problems. While Wikipedia aims to organise its categories as a directed acyclic graph of polyhierarchical subsumption relations [3], it was found that the graph is, in fact, a noisy pseudo-hierarchical structure with cycles [22]. Similarly, Boldi and Monti [20] describe the WCS graph as a sparse and noisy category pseudo-forest that contains cycles, duplications, errors and oversights, and that is often too fine-grained; therefore, the graph is difficult to use directly without prior modifications.

Despite these deficiencies, it has been shown that the WCS can be hierarchised by pinpointing a root and top-level categories and by organising the graph into hierarchy levels according to the length of directed paths in the graph [4,21]. Following this idea, we derived a hierarchical taxonomy from the categories of the English Wikipedia, making as few modifications to the original graph as possible, as described in the next paragraph. Given that only the first three to five levels of the WCS yield reasonably good subsumption relations [19], our taxonomy is limited to a depth of three levels. This limit is in line with our aim of evaluating a lean and manageable yet sufficiently detailed taxonomy for Use Case 2.

First, the 13 main categories from the English Wikipedia overview page were extracted. These categories are as follows: General Reference, Culture, Geography, Health, History, Human activities, Mathematics, Nature, People, Philosophy, Religion, Society and Technology. Some category names were modified slightly to make them more succinct (e.g. Health and fitness was shortened to Health). The categories Human activities and People were merged into a new category Person, since more than 90% of their articles had overlapped. The number of articles per category was retrieved with the MediaWiki API. A root node was added to connect the categories at the zeroth level of the graph. Subsequently, for each category, the subcategories along with the number of articles assigned to it were extracted. These numbers can be derived from Wikipedia dumps or using the MediaWiki API. For instance, the category Health has several subcategories such as Self-care, which by itself has more fine-grained subcategories like Fitness or Longevity. This process could be repeated several times for the subcategories, depending on the total depth of the respective branch. However, for the sake of manageability and the above-mentioned reasons identified in Bairi et al. [19], we extracted only the first three levels. Subcategories beyond the third level are very sparse.

At the second and third levels of the resulting graph, two problematic issues arise. First, there are instances of cyclic relations, meaning that some second-level subcategories are further categorised into one of the first-level categories. We removed links causing cycles. The second problem is that the graph contains duplicates of subcategories where a specific third-level subcategory has two different parents. We used frequency counts (i.e. how often a category is assigned to articles) of each subcategory to resolve ambiguous hierarchical relations.

The resulting three-level taxonomy is provided in the Supplementary Material.

4.2. Test corpus

To test whether the three-level taxonomy described in Section 4.1 is suitable for accurate and reliable labelling of open-domain text corpora, a crowd-based online annotation study was conducted. The volunteer annotators (see Section 4.4) were asked to use our designated online interface (see Section 4.3) to label the subject matter of eight short text fragments each. Broad thematic coverage and stylistic diversity of the test corpus was ensured by randomly sampling a total of 160 English texts from two sources: (1) the written component of the BNC XML Edition, a 90-million-word general-purpose corpus of British English covering a large variety of registers, genres, subject domains and linguistic styles; (2) the English Wikipedia, which is a ‘proxy for knowledge in general’ [4] (p. 2) due to its size and thematic diversity. The Wikipedia sample was created manually using Wikipedia’s random article selector, ignoring stubs shorter than three sentences. Based on BNC’s subject domain annotation, the thematic coverage of the BNC sample was assessed: the sample covers all nine BNC subject domains, with natural sciences and imaginative literature being underrepresented (Figure 1). The Wikipedia sample lacks unambiguous subject ground truth, because Wikipedia articles are usually assigned to multiple Wikipedia categories. We deem the test corpus a proxy to open-domain problems despite its limited size, because it combines samples from two large, thematically diverse and stylistically complementary representative text collections.

To avoid multi-topic texts and prevent demotivation of volunteer annotators, the sampled texts were truncated to approximately 120 words, respecting sentence boundaries. Such snippets are valid task stimuli, because usually the initial sentences of a text determine its ‘aboutness’ and establish the contact between writer and reader [23]. A breakdown of the test corpus is shown in Table 1.
4.3. Annotation interface

For the labelling of the test corpus, we developed a web-based interface that comprises three components: (1) instruction page with brief explanation of labelling procedure; (2) labelling panel; and (3) survey form. In the labelling panel, eight texts randomly selected from the test corpus were shown to annotators on separate pages, one page per text. Each text from the test corpus was assigned to three annotators. The overall study objective, as well as the taxonomy’s provenance, was not disclosed to avoid annotators’ bias. As shown in Figure 2(a), the annotators were tasked with labelling the subject matter of texts as precisely as possible by selecting appropriate entries from the taxonomy with the help of a three-level dropdown list. The dropdown list is a dependent one, that is, the options displayed at lower levels depend on the choices made on higher levels. Five-point rating scales next to each dropdown level captured how confident annotators were about their choices. Leaving any of the dropdowns or confidence ratings blank was not possible. Normally, in crowd-based subject labelling by non-experts annotators would be allowed to limit their label choices to the top level if they cannot find suitable labels at deeper levels of the taxonomy. However, in this study, annotators were required to choose labels at all three levels, because measuring labelling confidence is essential to assess the taxonomy’s intuitiveness. Furthermore, the labelling panel supported only a hierarchical use of Wikipedia categories. The reason for this is that we tested two more labelling modes (autocomplete-supported tagging with flat list of tags; combination of hierarchical taxonomy and free-text tagging) in a pilot study with an independent cohort of volunteer annotators, observing that the exclusive use of the hierarchical taxonomy yielded the most conclusive results. To ensure the labelling panel’s usability and to control for potential confounding variables, we decided to limit the panel to the hierarchical labelling mode only.
Figure 2(b) shows the survey form, which was displayed to annotators upon complete annotation of the eight texts. Through the survey form, annotators stated their opinion about the taxonomy by indicating their level of agreement with six qualitative statements on six-point Likert-type scales.

The collected data are provided in the Supplementary Material.

4.4. Annotators

Calls to voluntarily participate in the study were distributed through a number of academic mailing lists and social media channels in the field of (corpus and computational) linguistics, translation, natural language processing and digital
humanities, such as the LINGUIST List. The addressees were also encouraged to share the call with others. Participation in the study was open to anyone, not just to language experts or English native speakers. No remuneration was paid.

In total, 54 annotators volunteered in the study. Their self-reported profiles are summarised in Figure 3. The majority of annotators are academic language experts with advanced proficiency in English and thus representative of the target group of Use Case 2: researchers interested in well-labelled corpus data. The observed variation in annotators’ individual profiles approximates the crowd-based labelling scenario underlying this study. Annotators’ previous experience in corpus labelling was not assessed, because in real-life crowd-based labelling, it would be barely controllable.

5. Results and discussion

5.1. Descriptive analysis of the derived taxonomy

The derived monohierarchical taxonomy is a directed acyclic graph with 848 nodes, organised into 12 categories at the first level (L1) and 82 categories at the second level (L2). The average number of child nodes for first-level categories is 6.8 and for second-level categories 10.3. On average, each first-level category ends up in 70.7 leaves at the third level (L3). The average branching factor over the entire graph is 9.9. These numbers indicate increasing granularity at deeper levels. However, there is considerable variability in the granularity, especially at the second level, where the number of third-level child nodes ranges from 1 to 34 (M = 10.3, Mdn = 8, SD = 7.29). The three most fine-grained L2 categories are Religion > Religionism with 34 children at L3, Geography > Landforms with 32 children at L3 and Culture > Sports with 15 children at L3. By contrast, the L2 categories Culture > Recreation and Society > Government and Politics have only four and three L3 children, respectively. Consequently, labelling of texts concerned with the latter subject matters is inevitably less precise than for the fine-grained categories. These discrepancies in terms of granularity can be attributed to the unbalanced distribution of topics in Wikipedia [4,21,22]. As a matter of fact, even the fine-grained categories are inherently incomplete. For instance, the L2 category Culture > Sports has 15 child nodes referring to various types of sports; hence, the only way to label a text dealing with a sports topic not...
covered by these 15 categories would be to resort to the more coarse-grained L2 label *Sports*, at the expense of labelling precision.

Inconsistency and ambiguity, known to be the potential drawbacks of social tagging [13,8], can also be detected in our taxonomy. For instance, topics related to medicine appear at various positions of the graph: as the L2 category *Health > Medicine*, as the L2 category *Health > Human Medicine* or as the L3 category *Technology > Technology and Science > Medicine*. Such ambiguities may impede clear-cut labelling choices, thus reducing the reliability of the taxonomy as a controlled vocabulary for corpus labelling.

5.2. Validity of the taxonomy: IRA

Out of the 160 texts in the test corpus, 112 were labelled by three annotators each, while 48 texts were labelled by two annotators, thus yielding 432 total observations. From all labels available in the taxonomy, annotators used 12 different first-level labels (= 100% of all labels at L1), 61 different second-level labels (= 74.4%) and 190 different third-level labels (= 22.4%). The high percentages at L1 and L2 indicate broad topic coverage of the test corpus, which suggests that despite its limited size, the test corpus is representative of open-domain text collections. According to L1 labels (Figure 4), in the labelled data, humanities topics (e.g. Society, Culture, Person, History) are more prevalent than natural and applied sciences topics, with the notable exception of Technology. The distribution of L1 labels appears to mirror the composition of the test corpus, although the BNC subjects reported in Section 4.2 are not directly comparable to the WCS-based taxonomy labels. In Wikipedia, topics related to human sociocultural life are overrepresented [16,18], therefore the WCS-based taxonomy may have biased annotators towards labels from these categories.

To validate whether the three-level taxonomy is suitable for accurate hierarchical multi-class labelling of subject matters, the labels annotators assigned to texts were analysed in terms of IRA. Krippendorff’s alpha ($\alpha$) was used as the agreement statistic, because it treats the randomly sampled raters as freely permutable and interchangeable, thus not being biased by raters’ individual identities [24]. $\alpha$ was computed with the *irr* package [25] for R; 95% confidence intervals for $\alpha$ were computed with the standard bootstrap resampling method (10,000 iterations) recommended by Zapf et al. [26], as implemented in the *lamisc* R package [27]. We computed individual $\alpha$s for each level of the taxonomy, as well as overall $\alpha$s across all three levels. Overall $\alpha$s were computed in two different ways: first, by taking the mean of the three individual per-level $\alpha$s, and, second, by computing one multi-label $\alpha$ [28]. In the latter approach, agreement is not
computed on one label, but on one set of three labels. To account for partial agreement between annotators, for example, two annotators agreeing on the labels at L1 and L2 but disagreeing at L3, the Jaccard distance metric was used to capture partially overlapping multi-label sets. Overall multi-label $\alpha$ with Jaccard distance was computed with the NLTK toolkit [29].

Table 2 shows that overall multi-label $\alpha$ for all texts are only in the range of fair agreement, which by definition ranges from 0.21 to 0.40 [30]. Only at the first level we observe IRA in the moderate range (0.41–0.60 by definition), with $\alpha = 0.49$ (95% confidence interval: 0.41–0.56). The non-overlapping confidence intervals at the first two levels for all 160 texts show that the observed agreement difference is statistically significant at the 0.05 level. This was to be expected, because the first level of the taxonomy is much more coarse-grained than the remaining two levels, thus making labelling arguably easier at higher levels of the taxonomy. Yet, even at the first level, the taxonomy scores below the minimum of 0.667 required for acceptable agreement [31]. Taking into account all levels – irrespective of the method of overall agreement computation – the minimum acceptability of 0.667 is missed by a large margin. Therefore, we conclude that the derived taxonomy is not a sufficiently valid instrument for the manual labelling of the test corpus. At the third level, agreement is mostly only in the slight range (0.00–0.20 by definition), rendering the most fine-grained level of the taxonomy almost useless for the labelling task. While agreement is slightly higher for Wikipedia texts than for BNC texts (see Table 2), for both text sources, it is in the fair range only, except for the first level. The large overlaps of the confidence intervals at each level of Wikipedia and BNC texts suggest that the text source does not have an influence on IRA.

5.3. Completeness and intuitiveness of the taxonomy: annotator feedback

In this part of the study, we collate the findings on IRA from Section 5.2 with annotators’ feedback obtained through (1) the survey form displayed to annotators upon completion of the labelling task and (2) the five-point confidence rating scales for each level of the labelling panel.

In the survey form, annotators indicated their level of agreement with the following six qualitative statements:

- Q1: The texts I just categorised were easy to understand.
- Q2: The category system is logical.
- Q3: The category system should have more categories.
- Q4: The category system is easy to understand.
- Q5: The category system has too many categories.
- Q6: The category system should have more than three levels.

Q1 elicits data on perceived text comprehensibility; Q2 and Q4 on intuitiveness of the taxonomy; and Q3, Q5 and Q6 on its completeness. Responses to Q1 indicate that annotators’ understanding of the labelled texts was high: more than 75% of annotators fully or mostly agreed with Q1 (Figure 5). The high self-reported text comprehensibility suggests that the low IRA scores are likely not caused by overly difficult texts in the corpus. Furthermore, there was no significant association between annotators’ self-reported English proficiency and perceived text comprehensibility, as assessed by a
two-sided Fisher’s exact test ($p = 0.43$). Therefore, both annotators’ linguistic profiles and text comprehensibility can be ruled out as factors contributing to low IRA.

With regard to intuitiveness, more than 50% of annotators state full to slight agreement to Q4, but there is also a considerable percentage of disagreement (Figure 5). The picture is even less clear for Q2, for which we observe considerable variability in the responses. From these number we conclude that the intuitiveness of the taxonomy for untrained annotators cannot be taken for granted, although the responses to Q2 and Q4 are slightly skewed towards the positive range of agreement. This, in turn, means that variability in annotators’ intuitive understanding of and comfort with the taxonomy may have contributed to low IRA.

Finally, with regard to the taxonomy’s completeness, the depth of three levels was mostly deemed sufficient by annotators (Q6 in Figure 5). However, the number of categories (848) was mostly deemed insufficient: over two-thirds agree fully to slightly that the taxonomy should have more categories (Q3), while full to slight agreement to Q5 accounts only for a cumulative 18.5%. Annotators’ perceived incompleteness of the taxonomy can be interpreted in light of the heterogeneous granularity of the labels described in Section 5.1. In fact, one annotator proactively contacted us by email, reporting discomfort with the variability of granularity, which made it difficult to get used to the taxonomy on one hand and to precisely label texts on the other. We hypothesise that annotators’ high agreement to Q3 reflects the inherent incompleteness of static knowledge organisation systems when applied to open-domain problems.

Annotators’ self-rated labelling confidence (i.e. how sure they were about each chosen label) was captured through the five-point rating scales of the labelling panel. The data indicate that despite the completeness and intuitiveness issues identified above, annotators were mostly confident about the chosen labels, with a gradual decline from first-level to third-level labels (Figure 6). The distribution of confidence ratings differs significantly across levels, as assessed by a chi-square test ($p < 0.01$, $\chi^2 = 88.554$, df = 8). Hence, higher IRA at L1 (see Section 5.2) is also reflected in higher labelling confidence. However, annotators’ perceived confidence ratings for individual texts differ considerably, because IRA on the confidence ratings for all 160 texts is very low — in fact, even lower than IRA on the labels: $\alpha = 0.13$ for the first level (95% confidence interval: 0.02–0.24), $\alpha = 0.17$ for the second level (0.06–0.28) and $\alpha = 0.23$ for the third level (0.12–0.34). This means that while in absolute terms annotators reported relatively high labelling confidence, there is large variability as to which texts were labelled with high confidence. This, in turn, means that labelling confidence itself is not necessarily a predictor of labelling agreement.

Figure 5. Relative distribution of annotators’ responses to the statements Q1 to Q6 of survey form ($n = 54$).
To account for the individual differences in annotators’ confidence ratings, for each text, we computed mean confidence scores from the individual confidence ratings indicated by the annotators. Mean confidence scores were computed for all three levels of the taxonomy. A Kruskal–Wallis H-Test showed that there was a statistically highly significant difference in mean confidence ratings across labelling levels: \( \chi^2(2) = 47.689, p < 0.01 \), with mean confidence ratings of 4.2 for L1, 3.8 for L2 and 3.5 for L3. Post hoc pairwise comparison using Wilcoxon Rank Sum Test showed that mean confidence ratings at L1 differ significantly from L2 and L3, \( p < 0.01 \) each, while the difference in mean confidence ratings between L2 and L3 is not significant \( p = 0.053 \). This corroborates the results of the chi-square test computed on the raw confidence rating counts. In addition, it suggests that the granularity of labels has a significant impact on rating confidence: the more coarse-grained the label, the higher is raters’ perceived labelling confidence.

To analyse the effect of rating confidence on labelling agreement, the test corpus was divided into three distinct groups (high, medium and low labelling confidence) based on the mean confidence ratings for each text. Subsequently, IRA on the assigned labels was computed for each group. As shown in Table 3, on all three taxonomy levels, higher labelling confidence leads to higher labelling agreement. The non-overlapping 95% confidence intervals for texts labelled with high and medium confidence indicate that IRA for high-confidence texts is statistically significantly higher than for texts labelled with lower confidence. Yet, even at the coarse-grained first level, IRA for texts labelled with high confidence.
confidence is only in the moderate range and below the required minimum of 0.667. For low-confidence texts, the confidence intervals include zero and therefore indicate lack of agreement; however, the number of observations is very low and therefore the results are not robust. By means of a qualitative inspection of the labels assigned by annotators to texts in the low-confidence group, we identify the following factors that may explain low labelling confidence and thus low labelling agreement:

- Ambiguous categories: for example, Geography > Manmade features > Railway versus Society > Infrastructure > Rail transport;
- Fuzzy category boundaries: for example, Society > Education > Pedagogy versus Society > Education > Study skills;
- Ambiguous subject matters: texts may deal with multiple subject matters, so in a unary labelling task, it is difficult for annotators to decide which subject matter is the dominant one;
- Annotator-related factors: for example, misunderstanding of a category, misunderstanding of a text, lack of attention (such as choosing completely unrelated labels).

With regard to social tagging, the lack of a controlled vocabulary is seen as the major problem [8]. The first two factors listed above indicate, however, that in this labelling study, the structure of the controlled vocabulary itself (i.e. the Wikipedia-based taxonomy) may cause difficulties, too. The remaining two factors may be attributed to the lack of training of annotators in this study. However, it remains unclear whether annotator training and more detailed labelling guidelines may compensate for the apparent deficiencies of the taxonomy.

6. Conclusion and outlook

Numerous studies analyse the WCS for knowledge organisation (see Section 3). Evaluation is mostly carried out intrinsically, focusing on the structure of WCS or the knowledge organisation systems derived from it, for example, automatically induced taxonomies. Extrinsic evaluation using Wikipedia categories in actual indexing tasks is the exception [15,16]. This study contributes to the body of research on task-based evaluations of the WCS for knowledge organisation. It focuses on manual document labelling, which is relevant to various fields of application despite the widespread use of automatic indexing approaches, for example, for the creation of research-oriented text corpora or application-oriented gold-standard datasets.

The aim of this study was to evaluate whether the WCS reduced to a depth of three levels is a valid taxonomy for subject matter labelling of open-domain text corpora. The underlying use case is corpus labelling for browsing, searching and navigating text corpora, which is a prerequisite for numerous tasks in corpus and computational linguistics or related fields. Our results show that the derived three-level taxonomy is not a sufficiently valid instrument for the labelling of the test corpus which approximates open-domain labelling in this use case, because even at the most coarse-grained level, IRA is significantly below the acceptability minimum. Although the taxonomy mostly appears intuitive to annotators, it has deficiencies that limit its usefulness for manual open-domain labelling. Training of annotators in the use of the taxonomy would certainly improve results. However, from the data, we conclude that it is not annotators’ lack of training but the structure of the taxonomy itself that limits its usefulness for open-domain corpus labelling: the known drawbacks of social tagging [8,9] manifest themselves in our study, too. Even more severely, the reduction in the WCS to three levels makes it inherently incomplete, and the fallback option of using coarse-grained labels from higher levels of the taxonomy lacks accuracy, as evidenced by low agreement among annotators. A further problematic issue limiting the WCS’s usefulness for open-domain corpus labelling is that it aims at indexing encyclopaedia articles; therefore, applying it to other text genres may introduce bias.

So why is it that the WCS lacks usefulness for manual labelling tasks, although it was shown to be a viable resource for automatic approaches to knowledge organisation (see Section 3)? Automatic approaches can harness the entire category graph, while we derived a taxonomy of three levels only for the sake of manageability. Extending the taxonomy beyond three levels would make it unmanageable from a cognitive point of view, because increasing the depth of hierarchical lists results in considerably higher task complexity [32]. Open-domain corpus labelling that strikes a middle ground between completeness and precision on one hand and intuitiveness on the other remains a challenge. Several potential Wikipedia-based solutions could be explored in the future:

- Semi-automatic indexing: automatic recommendation of subject labels from a denoised taxonomy of Wikipedia categories [20], followed by manual selection of recommended labels;
Autocomplete-supported manual indexing using a flat list of Wikipedia categories: to mitigate the incompleteness problem of our three-level taxonomy, the entire (denoised) WCS graph might be used as the underlying controlled vocabulary;

Wikidata as controlled vocabulary: Wikidata, a collaboratively edited knowledge base of structured linked data [2,33–35], has recently received attention as a resource for semantic enrichment [33] and as a linked open data hub for library catalogues [34,35]. Wikidata items (similar to terms in controlled vocabularies) and properties (i.e. links between items) provide unique identifiers, labels designating preferred names, descriptions and aliases designating alternative names [34]. The items could be used as vocabulary terms for subject indexing and content retrieval, and the properties between Wikidata items would allow for browsing and navigation. The availability of labels and aliases for Wikidata items helps capture preferred and alternate terms (e.g. synonyms, spelling variants), thus promoting labelling consistency. In corpus labelling interfaces, Wikidata-supported labelling could be implemented in the form of autocomplete input boxes, which were proposed for intuitive semantic annotation based on linked open data resources [36]. Libraries increasingly recognise the benefits of using and contributing to Wikidata [34,35], thus making expert-curated vocabularies openly available through an easy-to-use repository.

Acknowledgements
The authors thank the annotators for voluntary participation in the study.

Declaration of conflicting interests
The author(s) declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding
The author(s) disclosed receipt of the following financial support for the research, authorship and/or publication of this article: This research is part of the project ‘TransBank: A Meta-Corpus for Translation Research’, funded by the goldigital 2.0 programme of the Austrian Academy of Sciences (grant number GD 2016/56).

ORCID iD
Michael Ustaszewski https://orcid.org/0000-0002-2000-5920

Data availability statement
The data underlying this study (extracted Wikipedia taxonomy, labelled test corpus, annotator feedback, annotator profiles), including licensing information, are provided as Supplementary Material.

Supplemental material
Supplemental material for this article is available online.

Notes
1. In this article, we distinguish between two often conflated terms: subject domain and subject matter. The former refers to a field or discipline that is interested in and/or studies a topic or set of phenomena, that is, the latter of the two terms. Subject matter labelling is more object-oriented than perspective-oriented: it focuses on what a text is about, that is, on its ‘aboutness’ [23], and not on who in particular deals with the respective topic. This focus increases the general usefulness of the labels. For instance, a legal text might be about organ transplantation, in which experts from both the legal and the medical domains might be interested. Labelling it as belonging to the legal domain would make it more difficult to find for medical experts, labelling it as having organ transplantation as its subject matter would make it easier to find for anyone interested in the topic.

2. In the literature, the term taxonomy is often used generically to denote any type of knowledge organisation systems, such as thesauri, subject headings, classification systems, controlled vocabularies, ontologies or folksonomies [37]; by contrast, for the sake of terminological clarity, we use taxonomies to denote taxonomies in the narrow sense, that is, hierarchically organised controlled vocabularies [37]. In the designation of the distinct types of knowledge organisation systems, we follow the definitions by Harping [13].

3. See www.natcorp.ox.ac.uk/archive/worldURG/urg.pdf

4. See https://en.wikipedia.org/wiki/Portal:Contents/Overviews

5. See https://www.mediawiki.org/wiki/API:Main_page
References

[1] Mehdi M, Okoli C, Mesgari M et al. Excavating the mother lode of human-generated text: a systematic review of research that uses the Wikipedia corpus. *Inf Process Manag* 2017; 53: 505–529.

[2] Nielsen FA. *Wikipedia research and tools: review and comments*. Lyngby: DTU Compute, 2019.

[3] Wikipedia. *Wikipedia: categorization*, 2020, https://en.wikipedia.org/wiki/ (accessed 25 May 2020).

[4] Suchecki K, Akdag Salah AA, Gao C et al. Evolution of Wikipedia’s category structure. *Adv Complex Syst* 2012; 15: 1–21.

[5] Tramullas J, Sánchez-Casabón AI and Garrido-Picazo P. Wikipedia categories in research: towards a qualitative review of uses and applications. In: Ribeiro F and Cerveira ME (eds) *Challenges and opportunities for knowledge organization in the digital age: proceedings of the fifteenth international ISKO conference*. Baden-Baden: Ergon, 2018, pp. 490–498.

[6] Greenberg J, Spurin K and Crystal A. Functionalities for automatic metadata generation applications: a survey of metadata experts’ opinions. *Int J Metadata Semant Ontol* 2006; 1: 3–20.

[7] Ustaszewski M and Stauber A. TransBank: Metadata as the missing link between NLP and traditional translation studies. In: *Proceedings of the workshop human-informed translation and interpreting technology* (eds I Temnikova, C Orasan, G Corspas, et al.), Varna, Bulgaria, September 2017, pp.29–25. Shoumen: ACL. DOI: 10.26615/978-954-452-042-7_004

[8] Peters I. *Folksonomies: indexing and retrieval in Web 2.0*. Berlin: De Gruyter, 2009.

[9] Trant J. Studying social tagging and folksonomy: a review and framework. *J Digit Inf* 2009; 10: 1–44.

[10] Thornton K and McDonald DW. Tagging Wikipedia. In: *Proceedings of the ACM 2012 international conference on support group work*. New York: ACM, 2012, pp. 219–227.

[11] Wu H, Zubair M and Maly K. Harvesting social knowledge from folksonomies. In: *Proceedings of the WikiSym ’06* (eds UK Wiil, PJ Nürnberg and J Rubart), Odense, 21–23 August 2006, pp. 111–114. New York: ACM.

[12] Fernando S, Hall M, Agirre E et al. Comparing taxonomies for organising collections of documents. In: *Proceedings of COLING 2012: technical papers* (eds B Kay and C Boitet), Mumbai, India, 8–15 December 2012, pp. 879–894. Mumbai, India: ACL.

[13] Harping P. *Introduction to controlled vocabularies*. Los Angeles, CA: Getty Publications, 2010.

[14] Kotlerman L, Avital Z, Dagan I et al. A support tool for deriving domain taxonomies from Wikipedia. In: *Proceedings of the international conference recent advances in natural language processing 2011* (eds R Mitkov and G Angelova), Hissar, 12–14 September 2011, pp. 503–508. Hissar: ACL.

[15] Ponzetto SP and Strube M. Taxonomy induction based on a collaboratively built knowledge repository. *Artif Intell* 2011; 175: 1737–1756.

[16] Josi F, Wartenla C and Charbonnier J. Text-based annotation of scientific images using Wikimedia categories. In: *Database and expert systems applications: proceedings of DEXA 2018 international workshops*, *BDMICS, BIOKDD, and TIR* (eds M Elloumi, M Granitzer, A Hameurlain et al.), Regensburg, 3–6 September 2018, pp. 243–253. Cham: Springer.

[17] Fernandes Medeiros J, Pereira Nunes B, Matsui Siqueira SW et al. TagTheWeb: using Wikipedia categories to automatically categorize resources on the web. In: *The semantic web: ESWC 2018 satellite events, revised selected papers* (eds A Gangemi, AL Gentile, AG Nuzzolese et al.), Heraklion, 3–7 June 2018, pp. 153–157. Cham: Springer.

[18] Voß J. Collaborative thesaurus tagging the Wikipedia way. *arXiv* 2006, cs/0604036v2, https://arxiv.org/ftp/cs/papers/0604/0604036.pdf

[19] Bairi RB, Carman M and Ramakrishnan G. On the evolution of Wikipedia: dynamics of categories and articles. In: *West R, Zia L and Leskovec J (eds) Wikipedia, a socialpedia: research challenges and opportunities – papers from the 2015 ICWSM workshop*. Palo Alto, CA: AAAI, 2015, pp. 6–10.

[20] Boldi P and Monti C. Cleansing Wikipedia categories using centrality. In: *WWW’16 companion: proceedings of the 25th international conference on World Wide Web* (eds J Bourdeau, JA Hendler, R Nkambou et al.), Montreal, QC, Canada, 11–15 April 2016, pp. 969–974. Geneva: International World Wide Web Conferences Steering Committee.

[21] Akdag Salah AA, Gao C, Suchecki K et al. Need to categorize: a comparative look at the categories of universal decimal classification system and Wikipedia. *Leonardo* 2012; 45: 84–85.

[22] Kittur A, Chi EH and Suh B. What’s in Wikipedia? In: *CHI ’09: proceedings of the SIGCHI conference on human factors in computing systems* (eds DR Olsen, RB Arthur, K Hinckley et al.), Boston MA, 4–9 April 2009, pp. 1509–1512. New York: ACM.

[23] Hutchins WJ. The concept of ‘aboutness’ in subject indexing. *Aslib Proc* 1978; 30(5): 172–181.

[24] Hayes A and Krippendorff K. Answering the call for a standard reliability measure for coding data. *Commun Methods Meas* 2007; 1: 77–89.

[25] Gamer M, Lemon J, Fellows I et al. irr: various coefficients of interrater reliability and agreement (R package version 0.84.1), 2019, https://cran.r-project.org/package=irr (accessed 15 April 2020).
[26] Zapf A, Castell S, Morawietz L et al. Measuring inter-rater reliability for nominal data — which coefficients and confidence intervals are appropriate? *BMC Med Res Methodol* 2016; 16: 93.

[27] Latour E. lamisc: Latour’s little helpers (R package version 0.0.0.9000), 2020, https://rdrr.io/github/emilelatour/lamisc (accessed 10 May 2020).

[28] Passonneau R. Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In: *Proceedings of the 5th international conference on language resources and evaluation* (eds N Calzolari, K Choukri, A Gangemi et al.), Genoa, 22–28 May 2006, pp. 831–836. Genoa: ELRA.

[29] Bird S, Klein E and Loper E. *Natural language processing with python*. Sebastopol, CA: O’Reilly, 2009.

[30] Landis JR and Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977; 33: 159–174.

[31] Krippendorff K. Content analysis: An introduction to its methodology. 3rd ed. Los Angeles: Sage, 2013.

[32] Jacko JA and Salvendy G. Hierarchical menu design: breadth, depth, and task complexity. *Percept Motor Skills* 1996; 82: 1187–1201.

[33] Mora-Cantallops M, Sánchez-Alonso S and García-Barriocanal E. A systematic literature review on Wikidata. *Data Technol Appl* 2019; 53/3: 250–268.

[34] ARL White Paper on Wikidata. *Opportunities and recommendations*. Association of Research Libraries, 2019, https://www.arl.org/resources/arl-whitepaper-on-wikidata

[35] Lemus-Rojas M and Pintscher L. Wikidata and libraries. In: Proffitt M (ed.) *Leveraging Wikipedia: connecting communities of knowledge*. Chicago, IL: ALA Editions, 2018, pp. 143–159.

[36] Trinh TD, Wetz P, Do B-L et al. An autocomplete input box for semantic annotation on the Web. In: *Proceedings of the international workshop on visualizations and user interfaces for ontologies and linked data (VOILA! 2015)* (eds V Ivanova, P Lambrix, S Lohmann et al.), Bethlehem, PA, 11 October 2015, pp. 97–102. Bethlehem: CEUR Workshop Proceedings.

[37] Hedden H. *The accidental taxonomist*. 2nd ed. Medford, OR: Information Today, 2016.