How Annotation Styles Influence Content and Preferences

Justin Cheng
Stanford University
Stanford, CA
jcccf@stanford.edu

Dan Cosley
Cornell University
Ithaca, NY
drc44@cornell.edu

ABSTRACT

Photo-tagging web sites provide several ways to annotate photographs. In this paper, we study how people use and respond to three different annotation styles: single-word tags, multi-word tags, and comments. We find significant differences in how annotation styles influence the objectivity, descriptiveness, and interestingness of annotations. Although single-word and multi-word tags are not normally differentiated, users prefer multi-word tags for their combination of descriptiveness and succinctness. We also discover that producers and consumers assess annotation styles differently in terms of ease of use, support for different user goals, and amount of effort required, demonstrating that allowing multiple modes of annotation is generally beneficial, as is considering both tag production and consumption.

Categories and Subject Descriptors

H.5.m. [**Information Interfaces and Presentation (e.g. HCI)**]: Miscellaneous

General Terms

Design; Experimentation.

Keywords

Tagging; annotation; digital photographs; motivations.

1. INTRODUCTION

Tagging, or more generally, annotation of digital objects has been an area of interest in recent years, with research done primarily in the *whys*, *whats* and *hows* of tagging, including the analysis of tagging vocabularies, indexing, and recommender systems for tags [22, 8, 7]. Building on existing understandings of annotation behavior, this paper focuses instead on annotation system design and its impact on users.

Motivations strongly influence how users tag objects. Although tags are often seen as tools for description that support the organization and retrieval of large amounts of data, tags in fact serve a number of expressive purposes. Ames and Naaman proposed a taxonomy that classifies user motivations for tagging images in terms of sociality (social vs. self) and function (organization vs. communication) [4]. For example, people may tag to help others find images (social/organization), or to help themselves remember events (self/communication). Within dimensions, motivations also vary widely: some users use tags to tell stories, or even make inside jokes [11]. As Nov et al. show through an analysis of user patterns and motivations within Flickr, a photo sharing website [12], understanding these characteristics is crucial for sustaining tagging communities.

With very different user motivations for annotation, designers are faced with the challenge of developing annotation systems that align user motivations with system goals like crowd-sourcing metadata or promoting discourse. Here, we explore how both producers and consumers of annotations use three annotation styles that might support different goals: single-word tags (SWTs) like "sunset" and comments like "That cactus looks like a fork!" that are common in social media systems, and a third style, multi-word tags (MWTs) such as "desert sky".

We believe that annotation styles may play a significant role in supporting and shaping a user's motivations, interpretations and uses of a system. For instance, Sen et al. [17] find that some MovieLens users apply tags that simply describe objects in images, while others add opinion and humor. Meanwhile, the ArtLinks project revealed that phrases were more interesting than SWTs, but SWTs were more descriptive and recurred more frequently, making them useful for connecting items [6]. Also, a study of Flickr data showed that tags were mostly related to places ("Italy") while other metadata like titles and captions were more narrative [10].

We build on this work in two ways. First, we further explore the lexical differences and uses of these annotation styles by producers. In particular, we are interested in MWTs. Though uncommon in social media, they fall between SWTs and comments, just as phrases come between words and sentences. Hence, we believe they may serve new expressive purposes and hit a "sweet spot" not served by SWTs and comments.

***RQ1**: How do the lexical and objective qualities of SWTs, MWTs and comments differ?*

Second, most of the work above focuses on the production of tags. However, annotation preferences may vary depending on whether a user is doing the annotations (a producer) or using others' annotations (a consumer). For instance, subjective, opinionated tags might be useful for self-expression for producers but of limited value for people searching for a certain kind of movie. Thus, we looked at how producers' and consumers' preferences for annotation styles may differ.

***RQ2a**: How do producers regard different annotation styles?*

***RQ2b**: How do consumers regard different annotation styles?*

We explore these questions with a two-part experiment in which we first asked 21 participants to annotate a common set of images using the three annotation styles, then asked 29 other participants to review and evaluate those annotations.

2. STUDY 1: TAGGING IMAGES

2.1 Method

In the first study, participants completed an online tagging task in which they were presented with 30 Flickr images, selected from a broad range of categories including events, people, places, and abstract art. Choosing a common set of images for people to tag is an approach that has also been used in other tagging studies [5, 18].

Participants were provided with brief explanations and examples of the 3 annotation styles, then instructed to annotate each image using a randomly chosen annotation style. The order of styles and images were randomized for each participant. The experiment was within-subjects and each participant tagged approximately one-third of the images using each style.

To control for interface differences between styles, we provided a single text box in each case, with minimal instructions indicating the type of annotation style the participant should use for that particular image (Figure 1). We did not show annotations made by previous users in order to avoid the influence of seeing others' annotations [17]. We also did not provide a specific context or goal for tagging, following many real systems like Flickr that do not instruct users how tags or comments should be used.

After the tagging task, participants completed a survey that asked about their prior tagging experience and their motivations for tagging, both in general and for this experiment. We also asked about their opinions with respect to ease of use, creativity, likelihood of online participation, and overall preference for each of the three annotation styles. As participants were tagging unfamiliar images, a majority reported that their main motivation for annotation was to *inform others about details of an image*. Thus, our results focus on the social, organizational, and communicative dimensions of tagging [4]; tagging for the general public was also shown to be the most common motivation for tagging on Flickr [13].

We chose to create our own interface and conduct a laboratory study using a common set of pictures in order to reduce the effect of how people might perceive a particular tagging system's uses and norms, as well as the effect of the interface design on on participants' motivations in producing or consuming annotation. Not only do annotation styles, layouts and input mechanisms differ in existing online systems—Flickr or 500px [1] allow both tags and comments, Tumblr [21] allows only tags (but not others to contribute tags), and Pinterest [15] allows only comments—-but the demographic and intended uses of tags and comments on these sites vary significantly. Further, MWTs (or phrase tagging) are not explicitly supported in any of these sites: users generally have to resort to unnatural workarounds like camel case or underscores, which may have further discouraged the use of such an annotation style in the wild.

2.2 Results and Discussion

21 undergraduates from 12 majors (15 female), recruited via a subject pool at a large Northeastern U.S. university, generated a total of 1275 annotations: 679 SWTs[1] (386 unique), 387 MWTs (353 unique) and 209 comments (all unique). Most had prior experience tagging people in images and commenting on Facebook, and a small proportion had used Flickr before. MWTs had a mean of 2.80 words, and comments had a mean of 6.17 words. Participant quotes from study 1 are labeled P#, while those from study 2 are labeled C#.

Figure 1: The annotation interface for MWTs. Similar interfaces were presented for both SWTs and comments.

***RQ1** How do the lexical and objective qualities of SWTs, MWTs and comments differ?*

2.2.1 Objectivity vs. Subjectivity

We defined objectivity as judgment relating directly to features of an image that most people would agree on, and subjectivity as based on opinion ("Mona Lisa is overrated"). Two coders independently coded each annotation as objective or subjective, with inter-coder disagreement resolved through discussion. Inter-coder agreement was 92.0% ($\kappa = 0.767$).

We found that 12% of SWTs were subjective, 25% of MWTs were subjective, and 48% of comments were subjective. That is, the proportion of subjective comments was almost double that of MWTs, which was in turn double that of SWTs. ($\chi^2 \geq 21.1$, $p < 10^{-5}$). While Sen et al. found that 24% of tags were subjective when users were asked to tag movies without seeing other users' tags [17], our results corroborate Marshall's [10], with participants using tags more for labeling and identifying an image and comments for more evaluative descriptions. As P12 put it, "I think commenting asks for my own opinion more than simply what I'm seeing." This difference may have resulted from the difference in available field formats—MovieLens only provided tags as a method of input, while our interface, and Marshall's, allowed both tags and comments.

We also used Linguistic Inquiry and Word Count (LIWC) [14] to look for differences in how annotation style included words from LIWC categories we thought would relate to motivations for annotation found in prior work [17, 10]. We selected categories that broadly related to the different possible motivations for tagging: to simply describe an image ("perceptual" words like "feels" or "view", as well as "relative" words relating to motion, space and time like "area" or "stop"), or to express one's opinion, whether through an emotional reaction ("affective" words relating to positive and negative emotion) or a subjective evaluation ("cognitive" words relating to insight and causation like "think" and "know").

Figure 2 summarizes these results. LIWC analysis at the word level revealed that comments had significantly ($p < 0.05$) more words categorized as cognitive than MWTs ($\chi^2(1,1760) = 4.79$), that MWTs had more perceptual words than comments ($\chi^2 = 5.44$), and that SWTs had less words that indicated relativity than MWTs or com-

[1]A small proportion (1%) of SWTs entered were made up of multiple words, which we discarded.

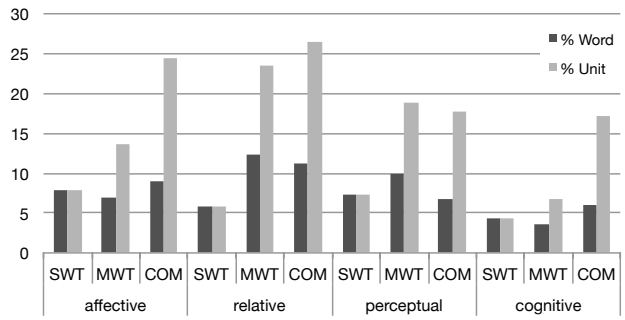

Figure 2: % of SWTs, MWTs or comments in each LIWC category. We measured category percentages both by breaking an annotation into its constituent words, and as a whole unit.

ments ($\chi^2(1,1529) = 18.3$, $\chi^2(1,1577) = 13.1$ respectively). In other words, comments expressed more of a person's thoughts and judgments, MWTs were more descriptive in terms of sensory perception, while SWTs were unlikely to indicate time, location or motion.

If we look at the level of the whole unit rather than breaking each down into words, we find that comments are more emotional than MWTs, which are in turn more emotional than SWTs ($\chi^2(2, 1341) = 48.1$). We also observe lower p-values ($<10^{-10}$) in the differences between annotation styles terms of cognition, relativity and sensory perception (χ^2=46.4, 95.6, 36.9).

***RQ2a**: How do producers regard different annotation styles?*

Table 1 shows that producers had no significant overall preference for any annotation style. However, combining preference rankings on specific attributes such as ease of use with participant comments, we observed that participants fell into two broad groups: those that preferred SWTs ("taggers"), and those that preferred comments ("commenters").

2.2.2 Ease of use

48% of producers found SWTs easiest to use (the taggers), while 38% found comments easiest (the commenters). Producers also treated MWTs more like tags than like comments: 80% of taggers ranked comments as harder than MWTs, while 38% of commenters ranked SWTs as harder than MWTs.

Taggers cited entry speed, saying that "SWTs were easy and quick to think of" (P1), contrasting tags with comments which were "harder to come up with" (P16) and "forced [one] to write more than [one] wanted" (P18). In contrast, commenters emphasized comments' expressiveness: "[SWTs] limited what I wanted to say...with comments I could get across the exact message I wanted" (P4).

In relation to MWTs, participants were divided. Some mentioned that MWTs, with their phrase-like quality, actually felt more "natural" because SWTs were "more limiting" (P21) and comments made them "feel a need to write more about things [they] did not have much to say [about]" (P10). Others found MWTs less natural, because it was "harder to find a few words that went together" (P6), preferring SWTs where they could "tag spontaneously" (P20) or comments where they need not "worry about [a] word limit" (P7).

Study	Style	Liked Most	Ok	Liked Least
Producers	SWTs	9	5	7
	MWTs	6	9	6
	Comments	6	7	8
Consumers	SWTs	8	15	6
	MWTs	12	8	9
	Comments	9	6	14

Table 1: Producers' and consumers' overall preferences for each annotation style. While producers are generally indifferent between styles, consumers seem to prefer SWTs and MWTs to comments.

2.2.3 Creativity

Producers saw comments as most creative, followed by MWTs. Comments and MWTs seem to support creativity in different ways: "[The] superfluity [of comments] encourages creativity. [The] simplicity [of MWTs] does also" (P5). While comments supported creativity through their freeform nature, several participants felt that MWTs were interesting because of their semi-rigidness: "[MWTs] made me think of new ways to say things...[with comments] too ambiguous to be creative towards," (P13) and "[MWTs] made you think the most of your word choice" (P7).

Thus, participants felt a tradeoff between simplicity and creativity. While SWTs are easy to use, they do not lend themselves to narratives and opinions. While comments are rich and varied, they also take more effort to enter into a system. In this aspect, MWTs may serve as a happy tradeoff, offering significantly more creativity with a little more effort on the user's part, although they may initially be unfamiliar to users.

3. STUDY 2: EVALUATING TAGS

3.1 Method

To understand how users would evaluate annotation generated using these 3 specific styles, participants performed a tag evaluation task, then completed a survey. Participants were shown the same 30 images from study 1. For each image, they saw all of the annotations generated by producers in study 1 using one of the annotation styles. This also helped control the conditions under which these annotations were generated. Annotation style for each image and image order were randomized, within subjects. Participants rated how accurate, searchable, and interesting the annotations were on a 5-point Likert scale, following Al-khalifa and Davis [3]. Finally, they completed a survey similar to that in the first study about their motivations and preferences with respect to image annotation.

3.2 Results and Discussion

***RQ2b**: How do consumers regard different annotation styles?*

A total of 29 undergraduates (22 female, 13 majors, with no participant overlap) evaluated the annotations applied by the students from the first study; their ratings are shown in Table 2. Consumers preferred MWTs overall, followed by SWTs, and finally comments. Table 2 presents the mean Likert scale ranking for each annotation style on each of 3 aspects of annotation: accuracy, searchability, and interestingness. To test for significant differences, we used a mixed model with MCMC sampling[2]. These rankings are summarized below.

[2]We fitted linear mixed effects models to our data, and then used Markov chain Monte Carlo sampling to determine each model's probability distribution.

Accuracy: MWTs ≻ SWTs ≻ Comments. MWTs performed best, and SWTs and MWTs were perceived as objective in describing an image, while comments were perceived as "personal", "irrelevant" (C1), "opinionated" and "contradictory" (C3). Many felt MWTs "provid[ed] more description" (C21) than SWTs, without the "unnecessary words" (C17) that came with comments.

Searchability: MWTs ∼ SWTs ≻ Comments. Information foraging suggests that keywords are more useful than chunks of text when scanning lots of documents [16], and we see this effect here too, with tagging (MWTs and SWTs) rated as more searchable than comments. Consumers saw MWTs as "[describing] different characteristics of the image" (C7) and having "more applicable words" (C3) than SWTs. Searchability and accuracy ratings were correlated, ($\rho = 0.547,\ p < 10^{-15}$), aligning with particpant comments that accurate annotations would also be searchable.

Interestingness: Comments ≻ MWTs ≻ SWTs. Comments shone in terms of interest and engagement, as the large variety and opinionated nature of comments provided "interesting interpretations" (C23) and "encouraged discussion" (C21). SWTs were seen as least interesting, simply "[stating] what the image is about" (C7).

4. DISCUSSION

Figure 3 summarizes our main results on a rough continuum. Users have different preferences for annotation styles depending on their role in the annotation generation process, and the different annotation styles themselves also provide tradeoffs between accuracy, simplicity, creativity, and effort. Below, we summarize the three main takeaways our work offers for designers and researchers of annotation systems.

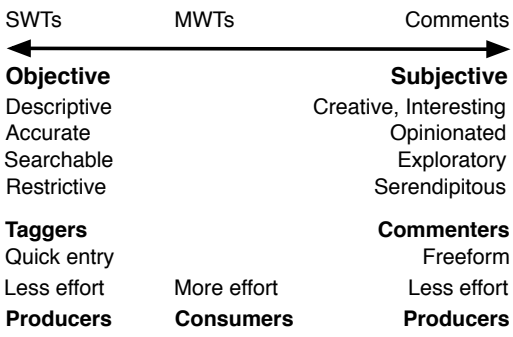

Figure 3: A continuum of perceptions of the 3 annotation styles.

4.1 Roles and effort matter

Style	Accurate	Searchable	Interesting
SWTs	3.98	3.62	3.00
(vs MWTs)	$(-^{\star})$	$(-)$	$(-^{\diamond})$
MWTs	4.09	3.88	3.25
(vs Comments)	$(+^{\ddagger})$	$(+^{\ddagger})$	$(-^{\diamond})$
Comments	3.73	3.37	3.32
(vs SWTs)	$(-^{\ddagger})$	$(-^{\dagger})$	$(+^{\ddagger})$

Table 2: Average Likert scale rating for different styles and comparison to other styles ($p < {}^{\diamond}0.1, {}^{\star}0.05, {}^{\dagger}0.01, {}^{\ddagger}0.001$). MWTs score well compared to other styles overall.

Consumers evaluate annotations differently from producers primarily because of the effort involved. While consumers preferred MWTs (12) to SWTs (8) and comments (9), their preferences shifted to SWTs (14, vs. 9 for MWTs and 6 for comments) when asked how likely they would be to use those styles as a producer, and these preferences were similar to those of producers. Consumers and producers also differed around their preferences for comments. Although a significant group of producers preferred commenting, 62% of consumers preferred commenting least, and when asked how likely they would be to create the different annotation types, both producers (50%) and consumers (62%) ranked comments as least preferable.

Sinha suggests tagging is easier than categorization because it is less cognitively demanding, and we suspect this might be the case here as well [19]. Consumers evaluated annotation in terms of how descriptive or interesting it would be, while producers balanced this with how easy it would be to generate these annotations[3]. Both producers and consumers also saw increasing text length as a barrier to contribution, which is especially apparent on mobile devices, which naturally lend themselves to systems like ZoneTag [2] and one-tap "likes".

Our results suggest that designers and researchers should pay more attention to the consumption aspects of annotation systems and to ways to reduce the effort of generating annotations.

4.2 Annotation styles support varying goals

Most consumers (86%) saw different annotation styles as supporting different goals. SWTs were associated with search, to "tag to help people find images" (C10, P18), and comments with discourse and reflection, "[wanting her] opinion to be known" (C10) and "[caring] more about people know what my pictures are about than if they can search for them" (C14). MWTs were called out as being like captions, offering rich description with conciseness (P16, C22).

In a similar vein, while short tags might be more efficient when quickly skimming documents or foraging for information, Lin et al. showed that sentences (and thus longer annotations like comments) are better at helping users comprehend images [9].

Thus, systems should generally support multiple annotation styles. In systems that provide only a single style like MovieLens and ArtLinks, tags are appropriated for many uses. Supporting multiple styles might increase expressiveness while encouraging compartmentalization of these uses. To better understand how tags and comments relate in real systems, it would be interesting to see whether the split between objectivity of tags and comments is more pronounced when both, rather than only one, are available at the same time. While MWTs occasionally contained interesting quips, would users restrict themselves to using comments exclusively for snarky one-liners if given the option to both tag and comment? Our initial observations of Flickr activity seem to suggest this: while people used tags primarily for description and categorization, they almost always preferred to express their opinions through comments.

4.3 The allure of MWTs

Producers and consumers liked MWTs for their "good balance between [SWTs] and commenting" (P6). Combining accuracy, searchability and interestingness, consumers rated MWTs highest, recognizing how they combined the succinctness of SWTs with the interestingness of comments. MWTs were seen as supporting both search and discourse, with participants calling them out as being

[3]One also wonders whether people enjoy generating opinions more than reading them.

more descriptive (C22, C23) than SWTs, but also more "succinct" (C18) in comparison to comments. MWTs force users to think in terms of phrases, leading them to generate either interesting "quips" like "circle of life" when describing a seagull with a starfish in its mouth, or concise descriptive phrases like "racecars driving on a track". Lexically, MWTs were similar to comments in terms of relative and perceptual words, and between SWTs and comments in affective and cognitive words. Both producers and consumers noted that MWTs were the most descriptive, although their phrasal nature made it difficult for some producers to generate. Thus, although MWTs are relatively uncommon in the wild, our results suggest they might be valuable additions to current annotation styles.

However, a design challenge remains in generating effective interfaces to support multi-word tagging. Sukumuran et al. proposed the use of situational norms, or showing users desired examples to encourage them to contribute similarly [20]; showing pre-generated MWTs could encourage new users to also tag in a similar fashion.

4.4 Limitations

The lab experimental setting allowed us to control the conditions under which annotations took place, to collect a corpus of annotations on the same set of pictures, to easily access and survey both producers and consumers of tags, and to explore MWTs. This approach complements prior work that analyzed tag corpora downloaded from real systems, but does exclude the personal motivations and natural context of annotation. Still, we believe it valuable to understand how users tag images that aren't necessarily their own, especially with the increasing prevalence of photo-centric web sites like Tumblr and Pinterest where users do not necessarily share their own personal photographs. Based on our work's concordance with prior work around the lexical style of tags, we expect that the perceived uses and lexical content of different annotation styles will be relatively stable across contexts and goals, and that producers will gravitate towards annotation styles that best fit their goals. That said, to understand the motivations and community processes involved in the use of each annotation style in the context of real systems and people's own content would be valuable future work.

5. CONCLUSION

In this paper, we studied how users created and evaluated different annotation styles for pictures, finding that they had differing preferences for annotation style, depending on whether they were producers or consumers. Each style elicited distinct reactions and was seen as useful for different purposes, suggesting that systems can vary or combine annotation styles to encourage particular user behaviors or design goals. In particular, MWTs appear to provide nice balance between the conciseness and searchability of SWTs, and the expressiveness of comments. Although MWTs are not offered as a distinct annotation style in current systems, our results suggest that making them easy to add and visible in the interface could lead to new goals and kinds of expressiveness compared to current tagging systems.

Followup experiments that use people's own data, analyze people's practices around annotations in real systems, and explore MWT interfaces in the field are in order to extend and refine these results. Still, our findings suggest that more attention to roles, goals, and innovative annotation tools will lead to more effective systems for both users and system owners.

6. ACKNOWLEDGMENTS

We would like to thank Karin Patzke and Evan Earle for their help in the experimental design and interpretation of results.

7. REFERENCES

[1] 500px. http://www.500px.com.

[2] Ahern, S., et al. ZoneTag: Designing Context-Aware Mobile Media Capture to Increase Participation. In *Workshop on Pervasive Image Capture and Sharing, UbiComp '06* (2006).

[3] Al-khalifa, H., and Davis, H. Measuring the Semantic Value of Folksonomies. In *Innovations in Information Technology*, IEEE (2006).

[4] Ames, M., and Naaman, M. Why we tag: motivations for annotation in mobile and online media. In *CHI '07* (2007), 971–980.

[5] Bischoff, K., Firan, C. S., Nejdl, W., and Paiu, R. Can all tags be used for search? In *CIKM '08* (2008), 193–202.

[6] Cosley, D., et al. Artlinks: fostering social awareness and reflection in museums. In *CHI '08* (2008), 403–412.

[7] Gemmell, J., Ramezani, M., Schimoler, T., Christiansen, L., and Mobasher, B. The impact of ambiguity and redundancy on tag recommendation in folksonomies. In *RecSys '09* (2009), 45–52.

[8] Kipp, M., and Campbell, D. G. Patterns and inconsistencies in collaborative tagging systems: An examination of tagging practices. In *ASIST'06* (2006).

[9] Lin, J., et al. What makes a good answer? the role of context in question answering. In *INTERACT '03* (2003), 25–32.

[10] Marshall, C. C. No bull, no spin: a comparison of tags with other forms of user metadata. In *JCDL '09* (2009), 241–250.

[11] Miller, A. D., and Edwards, W. K. Give and take: a study of consumer photo-sharing culture and practice. In *CHI '07* (2007), 347–356.

[12] Nov, O., Naaman, M., and Ye, C. What drives content tagging: the case of photos on flickr. In *CHI '08* (2008), 1097–1100.

[13] Nov, O., and Ye, C. Why do people tag?: motivations for photo tagging. *Commun. ACM 53*, 7 (July 2010), 128–131.

[14] Pennebaker, J., and Francis, M. Linguistic inquiry and word count: LIWC 2007. *Mahway: Lawrence* (2001).

[15] Pinterest. http://www.pinterest.com.

[16] Pirolli, P., and Card, S. Information foraging in information access environments. In *CHI '95* (1995), 51–58.

[17] Sen, S., et al. tagging, communities, vocabulary, evolution. In *CSCW '06* (2006), 181–190.

[18] Sigurbjörnsson, B., and van Zwol, R. Flickr tag recommendation based on collective knowledge. In *WWW '08* (2008), 327–336.

[19] Sinha, R. A cognitive analysis of tagging. `http://rashmisinha.com/2005/09/27/a-cognitive-analysis-of-tagging/`, 2005.

[20] Sukumaran, A., Vezich, S., McHugh, M., and Nass, C. Normative influences on thoughtful online participation. In *CHI '11* (2011), 3401–3410.

[21] Tumblr. http://www.tumblr.com.

[22] Zubiaga, A., Martínez, R., and Fresno, V. Analyzing tag distributions in folksonomies for resource classification. In *KSEM'11* (2011), 91–102.