Analysis of the impact of a tag recommendation system in a real-world folksonomy

Frederic Font*[1], Joan Serrà[2], and Xavier Serra[1]

[1]Universitat Pompeu Fabra
[2]Artificial Intelligence Research Institute (IIIA-CSIC)
[1]Universitat Pompeu Fabra

Abstract

Collaborative tagging systems have emerged as a successful solution for annotating contributed resources to online sharing platforms, facilitating searching, browsing, and organising their contents. To aid users in the annotation process, several tag recommendation methods have been proposed. It has been repeatedly hypothesized that these methods should contribute to improve annotation quality as well as to reduce the cost of the annotation process. It has been also hypothesized that these methods should contribute to the consolidation of the vocabulary of collaborative tagging systems. However, to date, no empirical and quantitative result supports these hypotheses. In this work, we deeply analyse the impact of a tag recommendation system in the folksonomy of Freesound, a real-world and large-scale online sound sharing platform. Our results suggest that tag recommendation effectively increases vocabulary sharing among users of the platform. Also, tag recommendation is shown to contribute to the convergence of the vocabulary as well as to a partial increase in the quality of annotations. However, according to our analysis the cost of the annotation process does not seem to be effectively reduced. Our work is relevant to increase our understanding about the nature of tag recommendation systems, and points to future directions for the further development of those systems and their analysis.

*This work has been supported by BES-2010-037309 FPI from the Spanish Ministry of Science and Innovation, TIN2009-14247-C02-01 from the Spanish Government (F.F.), 2009-SGR-1434 from Generalitat de Catalunya (J.S.), JAEDOC069/2010 from CSIC (J.S.), FSE2007-2013 E. U. Social funds (J.S.), and FP7-2007-2013 / ERC grant agreement 267583 (CompMusic; F.F. and X.S.). Author's addresses: F. Font, Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain; email: frederic.font@upf.edu; J. Serrà, Artificial Intelligence Research Institute (IIIA-CSIC), Spanish National Research Council, Bellaterra, Spain; email: jserra@iiia.csic.es; X. Serra, Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain; email: xavier.serra@upf.edu.

1 Introduction

Online sharing platforms make extensive use of semantically-meaningful textual labels, called tags, to describe and annotate its contents. The use of these tags provides a means for searching, browsing and organising the resources of the platform. Systems that provide the functionality for making these annotations are normally referred to as collaborative tagging systems. In collaborative tagging systems, users of the online platform have the responsibility of annotating the content. Every relation between a tag and a content resource performed by a user of the system can be identified as a tag application (Sen et al., 2006). We refer to the set of all distinct tags that are assigned to a particular resource as the *tagline* of the resource. The aggregate of all tag applications, which relate tags, resources and users of an online sharing platform, is normally known as the folksonomy (Vander Wal, 2007).

In general, tags introduced using collaborative tagging systems are not restricted in its form, and users can freely create new tags at any time (Marlow et al., 2006; Sen et al., 2006; Wagner et al., 2014). This provides a great flexibility to collaborative tagging systems as opposed to other systems that make use of pre-defined vocabularies, and in which users are not allowed to annotate content using terms that are not included in these vocabularies (Robu et al., 2009; Wagner et al., 2014). With non-restricted vocabularies, users introduce new tags when the annotation of a particular resource requires it. Hence, they easily adapt to the evolution of the platform's content. Furthermore, it has been suggested that users feel more comfortable during the annotation process when they are not restricted to the use of a pre-defined vocabulary (Robu et al., 2009).

Collaborative tagging systems suffer from a number of well-known problems including tag scarcity, the use of different tags to refer to a single concept (synonymy), the ambiguity in the meaning of certain tags (polysemy), typographical errors, the use of user-specific naming conventions, or the use of different languages (Halpin et al., 2006). It is often discussed whether the folksonomy of a collaborative tagging system, after a certain time of being in use, reaches a point of implicit consensus. In that point of consensus, the vocabulary is supposed to converge to a certain set of tags and tagging conventions that are widely adopted by all users of the system (Halpin et al., 2006; Sen et al., 2006; Sood et al., 2007; Robu et al., 2009; Wagner et al., 2014). Such a consensus implies more coherent resource annotations and better opportunities for searching, browsing, and organising content (Spiteri, 2013). Additionally, it leverages the value of the folksonomy as a source of knowledge mining (Wagner et al., 2014). Some studies have analysed this aspect, and the emergence of a consensus has been highlighted in several occasions (Robu et al., 2009; Wagner et al., 2014). According to these studies, the emergence of consensus depends on several factors, one of them being

the way in which users are exposed to the annotations performed by other users. In general, the more users are exposed to the tagging conventions of other users, the fastest should the consensus emerge.

In order to try to overcome some of the issues of collaborative tagging systems, tag recommendation systems can be employed to suggest potentially relevant tags during the annotation process of a resource (Jäschke et al., 2007). These systems are generally based on the analysis of the content of the resources being annotated, or in the folksonomy of a collaborative tagging system. Former systems normally use feature extraction techniques to analyse content resources, and further training of machine learning models that can predict tags based on the extracted features (e.g., Li and Wang 2008, Turnbull et al. 2008, Toderici et al. 2010). Folksonomy-based systems normally take advantage of tag co-occurrence information in previously annotated resources in order to provide relevant tag recommendations for newly annotated resources (e.g., Sigurbjörnsson and Zwol 2008, Garg and Weber 2008, De Meo et al. 2009, Ivanov et al. 2010, Font et al. 2013b).

It can be intuitively hypothesized that a tag recommendation system, independently of its nature, should have an impact on the folksonomy of a collaborative tagging system. In fact, this has been suggested by many authors. Golder and Huberman 2006 hypothesize that a tag recommendation system should help consolidating the tag vocabulary across users. The same idea is suggested by Jäschke et al. 2007 and Marlow et. al. 2006. Jäschke et al. 2007; 2012 also hypothesize that tag recommendation should simplify the process of finding good tags for the resources being described and thus increases the chances of getting resources annotated. Similarly, Sood et al. 2007 hypothesize that by using a tag recommendation system, users can see how other users tag resources and better choose when to reuse already existing tags or when to create new ones. Therefore, tag recommendation should help alleviate synonymy problems and help vocabulary convergence (Sood et al., 2007). These authors also hypothesize that the use of a tag recommendation system fundamentally changes the tagging process from being a generation process, where users must create tags from scratch, to being a recognition process, where users have to recognise valid tags from a list of suggestions. Zangerle et al. 2011 perform a study on *hashtag* recommendation for Twitter[1], a microblogging site, and hypothesize that the use of hashtag recommendation should help homogenising hashtags. Finally, Wang et al. 2012 hypothesize that tag recommendation can improve both the quality of tags and the efficiency of the tagging process, by clarifying the semantics of tags and reducing the manual cost of tagging.

Taking into consideration the previous statements, we can summarise the expected impact of a tag recommendation in the folksonomy of a collaborative tagging system in the following three hypotheses:

[1] `http://www.twitter.com`

1. *Vocabulary convergence.* A tag recommendation system should contribute to the convergence and consolidation of a shared vocabulary across the users of a collaborative tagging system (Golder and Huberman, 2006; Marlow et al., 2006; Jäschke et al., 2007; Sood et al., 2007; Zangerle et al., 2011).

2. *Quality of annotations.* A tag recommendation system should improve the quality of annotations of the resources in an online sharing platform (Jäschke et al., 2012; Wang et al., 2012).

3. *Cost of the annotation process.* A tag recommendation system should reduce the cost of tagging, changing from a tag generation process to a tag recognition process (Sood et al., 2007; Jäschke et al., 2007; Wang et al., 2012).

As mentioned, there have been many studies proposing different tag recommendation methods. Some of them evaluate the quality of the recommendations using data from real-world folksonomies (e.g., Sigurbjörnsson and Zwol 2008, Jäschke et al. 2009, De Meo et al. 2009, Font et al. 2014b). Other studies are focused on analysing the characteristics of collaborative tagging systems (e.g., Marlow et. al. 2006, Halpin et. al. 2006, Golder and Huberman 2006, Farooq et. al. 2007, De Meo et. al. 2013). Nevertheless, we are not aware of any study performing a deep analysis of the impact of a tag recommendation system into a real-world and large-scale folksonomy. Thus, the three previous hypotheses remain unverified and lack empirical evidence.

In this work, we analyse the impact of a tag recommendation system into the folksonomy of Freesound, a sound sharing site with more than 3.7 million registered users and 200,000 uploaded sounds (Font et al., 2013a). In Freesound, users upload sounds and then annotate them, yielding a *narrow* folksonomy in which only the authors of the sounds can annotate them (Vander Wal, 2005). The Freesound folksonomy features 1.5 million tag applications involving 70,000 distinct tags and 10,000 different users (i.e., only a small fraction of registered users do upload sounds and thus generate tag applications). In 2013, eight years after Freesound was started, a tag recommendation system was introduced. That tag recommendation system is a folksonomy-based system described in previous work by the authors (Font et al., 2013b, 2014a,b). Here, we analyse the impact that this system has had in the folksonomy of Freesound. For each one of the three hypotheses that we summarised above, we define a series of metrics to illustrate them. Then, we compute these metrics for an extensive period of time comprising 2.5 years of analysis data, and analyse the results putting special emphasis on the changes observed before and after the introduction of tag recommendation. Our results give, for the first time, empirical and quantitative evidence of the validity of some of the previous hypotheses. Specifically, our

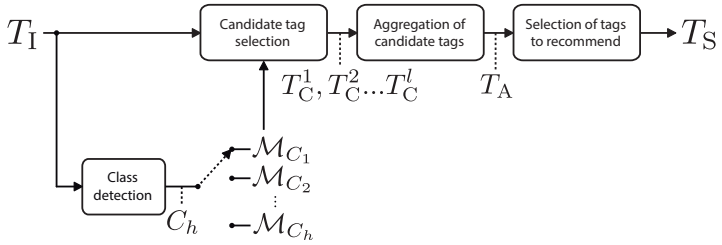

Figure 1: Block diagram of the tag recommendation system implemented in Freesound.

results show that the tag recommendation system effectively contributes to the vocabulary convergence of the folksonomy, partially contributes to an improvement of annotation quality, but does not seem to significantly reduce the cost of the annotation process. Despite our evaluation methodology is only applied to analyse the impact of a tag recommendation system in the context of Freesound, we believe that our results are indicative of the impact that tag recommendation systems can potentially have in other collaborative tagging systems. In closing, some suggestions are made regarding how could our analysis be extended, and tag recommendation systems be improved to further increase the impact on some of the analysed aspects such as the quality of the annotations. Both the definition of the metrics and the analysis of its results are also relevant contributions of the present work.

The rest of the paper is organised as follows. In Sec. 2, we briefly describe the implemented tag recommendation system and define the proposed evaluation metrics and analysis methodology. The results for all evaluated metrics, along with discussions about their implications, are reported in Sec. 3. In Sec. 4 we summarise our findings and discuss about the limitations of our analysis. We end this work in Sec. 5 by drawing some possible future directions.

2 Methodology

2.1 Tag recommendation system

The tag recommendation system implemented in Freesound is based on tag-tag similarity matrices derived from the folksonomy of the same platform. In this section, we briefly describe its main components. Exhaustive description and evaluation of the system can be found elsewhere (Font et al., 2013b, 2014a,b).

Given a set of input tags T_I, the system is able to output a set of recommended tags T_S. With that purpose, the system performs the following steps (Fig. 1):

1. Class detection: The first step consists in the classification of the input tags T_{I} into a set of H predefined audio classes. We defined $H = 5$ audio classes (*SoundFX*, *Soundscape*, *Sample*, *Music* and *Speech*), and built a ground truth by manually annotating 1,200 Freesound sounds per class (Font et al., 2014a). Using this ground truth, we trained a multivariate Bernoulli naive Bayes classifier, feeding it with the taglines of the sounds. Then, given a set of input tags T_{I}, the classifier can predict which category C_h better fits the input. Accuracies range between 75 and 95%, depending on the length of T_{I}.

2. Candidate tag selection: Given the set of input tags T_{I}, this step selects a pool of candidate tags T_{C}^{l} for each input tag T_{I_l}. We do so by choosing the top 100 most similar tags according to a tag-tag similarity matrix $\mathcal{M}_{C_h}$, which depends on the predicted class C_h of the previous step. Matrices $\mathcal{M}_{C_h}$ are computed offline and considering a model of the folksonomy of Freesound $\mathcal{F}$, which is represented as a tripartite hypergraph $\mathcal{G}(\mathcal{F}) = \langle V, E \rangle$ (Mika, 2007; Font et al., 2013b). In this model, vertices are given by three finite sets of objects, $V = U \cup T \cup R$ (users, tags and resources, respectively), and each edge $E = \{\{u, t, r\} | (u, t, r) \in \mathcal{F}\}$ represents a tag application, embedding the relation between a tag t, a resource r (a sound), and the user u that performed that tag application. Given $\mathcal{G}$, we derive a sparse association matrix $\mathcal{D} = \{\mathcal{D}_{i,j}\}$, $i = 1, \ldots |R|$, $j = 1, \ldots |T|$, which represents the associations between the $|R|$ sounds and the $|T|$ distinct tags available in Freesound ($d_{i,j} = 1$ if sound r_i is labeled with tag t_j, and $d_{i,j} = 0$ otherwise). We use the same classifier used in the class detection step to predict the class of all sounds in the association matrix given their tag applications. Then, given $\mathcal{D}$ and the list of sounds we predicted for every class C_h, we can compute the different tag-tag similarity matrices by filtering out all columns from $\mathcal{D}$ corresponding to sounds which do not belong to a particular class and then performing a matrix multiplication so that $\mathcal{M}_{C_h} = \mathcal{D}\mathcal{D}'$ ($'$ indicates matrix transposition). Applying a simple normalisation to the elements of $\mathcal{M}_{C_h}$, we obtain a matrix whose elements $\{\mathcal{M}_{t_i,t_j}\}$ correspond to the cosine similarity between tags t_i and t_j on the context of a particular audio class C_h (Font et al., 2013b, 2014b).

3. Aggregation of candidate tags: Given the sets T_{C}^{l} from the first step, candidates are assigned a score ε and aggregated into a single list of tags with scores T_{A}. Such score is determined by the candidate similarity-based ranking so that $\varepsilon = 1$ for the most dissimilar candidate to a given input tag and $\varepsilon = N$ for the most similar one. The scores of tags that are present in different sets of candidates T_{C}^{l} are added when aggregated to the final set T_{A} (Font et al., 2013b).

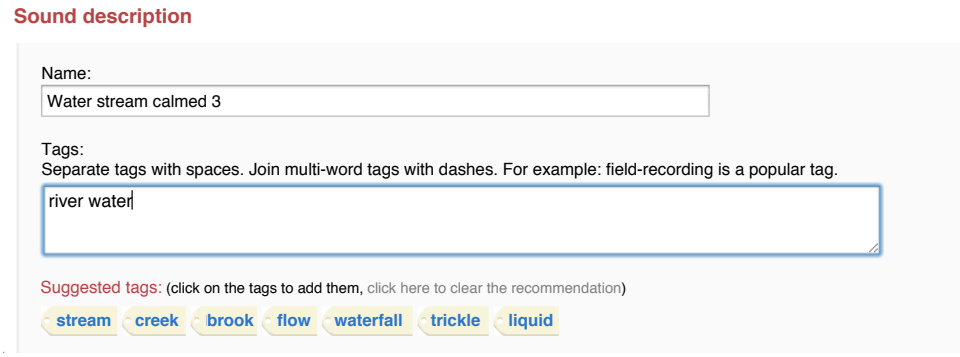

Figure 2: Screenshot of the interface of the tag recommendation system implemented in Freesound. The interface used in Freesound before the introduction of the tag recommendation system was exactly the same without the list of tag suggestions at the bottom.

4. Selection of tags to recommend: Considering the scores in T_{A}, this step determines a threshold ϵ to select the tags that are finally recommended. The threshold ϵ is set to be the 85% of the maximum score in T_{A}. Tags in T_{A} are sorted by their score and those that satisfy $\varepsilon \geq \epsilon$ are outputted as T_{S}, the final set of recommended tags (Font et al., 2013b).

2.2 Tag recommendation interface

Fig. 2 shows a screenshot of the interface for the tag recommendation system implemented in Freesound. In it, we can see the set of input tags T_{I} ={`river`, `water`} and the set of suggested tags T_{S} ={`stream`, `creek`, `brook`, `flow`, `liquid`, `waterfall`, `trickle`}. The list of suggested tags appears at the bottom of the text area that users use to type their tags. On average, the recommendation system produces lists of 4 suggested tags (Font et al., 2014b). However, if the recommendation system produces a list with more than 30 tags ($|T_{\mathrm{S}}| \geq 30$), only the first 30 are shown in the interface. This list of suggestions is automatically refreshed each time that users type a new tag (i.e., every time that T_{I} changes). This means that during the annotation process of a particular sound, several lists of suggested tags can be presented to the user. To introduce tags from the list of suggestions, users can either click on the elements of the list or type them manually as they would do to introduce tags that are not in the list. Freesound does not provide any kind of autocomplete functionality when manually typing tags.

Proposed metrics and expected observations to evaluate hypotheses. For the case the tag frequency distribution metric, we expect it to be more evenly distributed across the frequency range after the introduction of tag recommendation, specially reinforcing agreement on tags with less frequency.

Hypothesis	**Metric**	**Expectation**
Vocabulary convergence	Percentage of new tags	Decrease
	Average user vocabulary size	Increase
	User vocabulary sharing	Increase
	Sound vocabulary sharing	Increase
Quality of annotations	Average tagline length	Increase
	Percentage of misspelled tag applications	Decrease
	Tag frequency distribution	Even (see table caption)
	Subjective annotation quality	Increase
Cost of the annotation process	Average tag application time	Decrease
	Average percentage of correctly predicted tags	Similar to (Font et al., 2014b

2.3 Analysis metrics

To assess the impact that the tag recommendation system has on the folksonomy of Freesound, we define a series of metrics which are meant to illustrate the three hypotheses presented in Sec. 1. We illustrate each hypothesis with more than one metric, as we believe the relevance of the analysis particularly remains on the observation of changes simultaneously happening in several metrics, rather than the observation of a single metric being affected after the introduction of the tag recommendation system. Table 2.3 shows a list of the defined metrics, along with the changes we expect to observe when comparing data before and after the introduction of the tag recommendation system. Formal metric definitions subsequently follow, grouped by hypothesis.

2.3.1 Vocabulary convergence

- *Percentage of new tags*: This metric represents the percentage of tag applications performed during a given day of our analysis period which involve tags that were never used before in the folksonomy (i.e., tag applications that introduce previously non-existing tags in the folksonomy). Thus, this metric is computed on a daily basis (see Sec. 2.4). Considering the folksonomy model defined in Sec. 2.1, the percentage of new tags can be defined as

$$\eta_n = \frac{|T_n^{\text{new}}|}{|E_n|} \cdot 100,$$

where T_n^{new} is the set of tags that appeared for the first time in the n-th day of our analysis data, and E_n is the set of all tag applications performed during that same day (note that T_n^{new} cannot contain

duplicates, i.e., a particular tag cannot be considered as being "new" more than once). High values of η indicate that many new tags are being created and that, therefore, the vocabulary is not converging to a finite set of terms. Our expectation for this metric is that it should be reduced after the introduction of tag recommendation, as users will tend to reuse tags from the list of suggestions rather than creating new ones.

- *Average user vocabulary size*: This metric is also computed on a daily basis, and we define it as the total number of tag applications involving distinct tags that a user performed during a given day (i.e., the number of unique tags that a user assigned during a given day). Considering the folksonomy model defined in Sec. 2.1, the average vocabulary size can be expressed as

$$\varsigma_n = \frac{1}{|U_n|} \sum_{u \in U_n} |E_n^u|,$$

where E_n^u is the set of tag applications involving distinct tags that user u has performed during the n-th day of our analysis data, and U_n is the set of users that performed at least one tag application during that same day. High values of ς indicate that users employ a wide variety of tags for annotating their sounds, whereas low values indicate that users tend to employ always the same tags they have already used before. We believe that, using the tag recommendation system, users will be exposed to a wider variety of tags than the ones they would have thought of. Hence, we expect to observe a ς increase after the introduction of tag recommendation.

- *User vocabulary sharing*: This metric quantifies to which extent users employ tags that have also been employed by other users. To analyse this aspect we build a weighted network $\mathcal{U}$ where nodes represent users and edges represent the amount of tags shared between two users. Edge weights w between nodes i and j of $\mathcal{U}$ are normalised using standard Jaccard similarity. Given an arbitrary period of time k for which a network $\mathcal{U}_k$ can be constructed, the weight between two nodes can be computed as

$$w_{ij} = \frac{\left|T_k^i \cap T_k^j\right|}{\left|T_k^i \cup T_k^j\right|},$$

where T_k^i is the set of distinct tags that the user corresponding to the i-th node has annotated during the time period comprised in k (similarly for T_k^j and node j). In such a network, two users will be strongly connected if they use the same tags when annotating their sounds. Notice that, according to the definition above, every node in

$\mathcal{U}_k$ has a self-loop, i.e., for $i = j$ we have $w_{i,j} = 1$. Having defined $\mathcal{U}_k$, node strength (Barrat et al., 2004) acts as a basic indicator of the level of vocabulary sharing across users. The more strength the nodes have, the more tags users are sharing. Let L be the total number of nodes in $\mathcal{U}_k$, and ϑ_i be the node strength for the i-th node of $\mathcal{U}_k$ such that

$$\vartheta_i = \sum_{j=1}^{L} w_{ij},$$

we define user vocabulary sharing as the average node strength over the network so that

$$\mu(\mathcal{U}_k) = \frac{1}{L} \sum_{i=1}^{L} \vartheta_i.$$

In our analysis, we build two networks $\mathcal{U}_k$ as defined above, one considering all the data after the introduction of tag recommendation and the other considering data from a reference time window before the introduction of tag recommendation (see below). We compare these two networks by computing the difference between user vocabulary sharing (average node strength) in both networks. We asses the statistical significance of that comparison by taking the series of node strengths of both networks (i.e., without computing the average) and using the Kolmogorov-Smirnov two-sample test (Corder and Foreman, 2009) for evaluating the null hypothesis that the two samples of node strengths belong to the same distribution (we use a significance level of $p = 0.01$). After the introduction of tag recommendation, we expect to observe an increase in μ, as users will be highly exposed to the influence of tags used by other users, and therefore more links will be created in $\mathcal{U}$.

- *Sound vocabulary sharing*: Similar to the previous metric, we can also study the vocabulary sharing across sounds instead of users. In this way, sound vocabulary sharing represents the tags that sounds have in common. To analyse sound vocabulary sharing we build a weighted network $\mathcal{S}$ where nodes represent sounds and edges represent the number of tags that are common to the pairs of sounds linked by them. As in $\mathcal{U}$, edge weights are normalised using the Jaccard similarity, so that the weight w between nodes i and j of a network $\mathcal{S}_k$ computed from data for a time period k can be defined as

$$w_{ij} = \frac{\left|T^i \cap T^j\right|}{\left|T^i \cup T^j\right|},$$

where T^i is the set of tags assigned to the sound represented by the i-th node (similarly for T^j and node j). Notice that, in this case,

the definition of w_{ij} does not include the time period k in any of its terms. This is because all tag-tag applications for a given sound are done at once. Therefore, if the sound was uploaded in the time period k (and thus is represented by a node in the network $\mathcal{S}_k$), all its tag applications will have also been performed during that time period k. In $\mathcal{S}_k$, two sounds will be strongly connected if they are annotated with the same tags, and we consider node strength as a basic indicator of the vocabulary sharing across sounds. Thus, we can define sound vocabulary sharing ν for a network $\mathcal{S}_k$ as the average node strength over that network, and compute it in the same way as described for user vocabulary sharing.

For analysis purposes, we again build two networks with data before and after the introduction of tag recommendation. The two networks are compared in terms of their node strength following the same process described above for analysing user vocabulary sharing. After the introduction of tag recommendation, we expect to observe a ν increase, as users will be highly exposed to the influence of tags used by other users. Therefore, sound annotations will include these tags and more links will be created in the network $\mathcal{S}$.

2.3.2 Quality of annotations

- *Average tagline length*: This metric is computed on a daily basis, and we define it as the average number of tags assigned to sounds that have been uploaded during a given day. Considering the folksonomy model defined in Sec. 2.1, the average tagline length can be expressed as

$$\tau_n = \frac{1}{|R_n|} \sum_{r \in R_n} |E^r|,$$

where E^r is the set of tag applications involving a resource r and R_n is the set of sounds uploaded and annotated during the n-th day of our analysis data. High values of τ_n indicate that sounds are being annotated with many tags, with potentially more comprehensive descriptions. Our expectation for this metric is to observe an increase after the introduction of tag recommendation, as the provided list of recommendations will help users to add more tags during the annotation process. In fact, even if recommendations are not correct, they may serve as a guide for users, and convey which kinds of information should be annotated about the sounds being described. For instance, the recommendation system could suggest a tag like `120bpm` to a sound sample corresponding to a music loop of different tempo. However, this tag might suggest the user to describe tempo information and in this way generate a longer tagline (Font et al., 2014b).

- *Percentage of misspelled tag applications*: This metric represents the percentage of tag applications performed during a given day of our analysis period that contain tags with misspellings or typographical errors. Considering the folksonomy model defined in Sec. 2.1, the percentage of misspelled tag applications can be defined as

$$\omega_n = \frac{|E_n^{\text{miss}}|}{|E_n|} \cdot 100,$$

where E_n is the set of all tag applications performed during the n-th day of our analysis data, and E_n^{miss} is the set of tag applications performed during that same day which involve tags with misspellings. In order to estimate E_n^{miss}, we use a simple approach in which we check, for each individual tag, whether it exists or not in an English dictionary[2] (similarly to Guy and Tonkin 2006). We consider that these tags which do not appear in the English dictionary contain misspellings or typographical errors. Using such a simple approach, tags consisting of proper nouns, compound words, or written in other languages, are most likely considered to be misspellings. However, we assume that the presence of these kind of tags is not affected by the introduction of the tag recommendation system and thus our defined metric is meaningful enough for comparison purposes. High values of ω indicate that many of the tags assigned to sounds contain misspellings. Our expectation for this metric is that it should be reduced after the introduction of tag recommendation, as users will manually type fewer tags and choose them from the list of recommendations instead.

- *Tag frequency distribution*: One useful indicator of the impact of the tag recommendation system is the observation of changes in the frequency distribution of existing tags. Intuitively, tags that are very popular (i.e., that have high frequency) tend to correspond to broader semantic concepts, while less popular tags usually correspond to narrower ones. Looking at the tag frequency distribution we can thus have an idea of users' tagging behaviour and observe if it is influenced by the tag recommendation system. To do that, we compute the frequency of tags over a period of time k such that the frequency υ of a tag t can be expressed as

$$\upsilon_{t,k} = |E_k^t|,$$

where E_k^t is the set of all tag applications involving tag t during the time period k. We consider two time periods, one with data before

[2]For that purpose we use the open-source Enchant spellchecking library, with British English and American English dictionaries (`http://www.abisource.com/projects/enchant/`).

the introduction of tag recommendation and the other with data after tag recommendation, and compute the complementary cumulative distribution of tag frequencies over the two periods. These kind of plots are common within the collaborative tagging literature (Bischoff et al., 2008; Robu et al., 2009), and indicate the probability that the number of occurrences of a particular tag is above a certain level. By qualitatively comparing the resulting distribution over two periods of time, we can have an idea of in which tag frequency ranges the tag recommendation system has a bigger impact. Our expectation for this metric is that the tag recommendation system will make the distribution more even by reinforcing the usage of tags with less frequency.

Additionally, we compare the distribution of tag frequencies before and after the introduction of tag recommendation in terms of their fit into a power law distribution. It has been suggested that folksonomies whose distribution of tag frequencies can be fitted by a power law, exhibit mature vocabularies that lead to better quality descriptions (Mathes, 2004; Cattuto, 2006; Halpin et al., 2006; Wagner et al., 2014). Hence, we check if we observe any difference regarding this matter after the introduction of tag recommendation. This analysis is also directly related with the hypothesis that tag recommendation should contribute to the convergence and consolidation of the vocabulary of the folksonomy.

- *Subjective annotation quality*: We are interested in analysing whether the tag recommendation system has an impact on the quality of the annotations. To avoid having to define an absolute metric for quality, we opt for measuring quality in relative terms, by comparing the subjective quality of a set of annotations before and after the introduction of tag recommendation. To do so, we set up a small online experiment where participants were presented with pairs of sounds from Freesound along with their taglines, and had to judge which sound was, in their opinion, better annotated. Every pair of sounds consisted of one sound uploaded after the introduction of tag recommendation and another sound uploaded before that. Sounds were labeled as "Sound A" and "Sound B", without providing any links to the original sounds in Freesound and without giving any hint of which sound was uploaded before and after the introduction of tag recommendation (Fig. 3). For every participant, sound pairs were presented in random order, and the assignment of each sound as being "Sound A" or "Sound B", was also randomised. For every pair of sounds, participants could either answer that "Sound A" was better annotated than "Sound B", that "Sound B" was better annotated than "Sound A", or indicate that they did not think that one sound was better annotated than the other ("No preference"). If participants wanted to give further explanations for

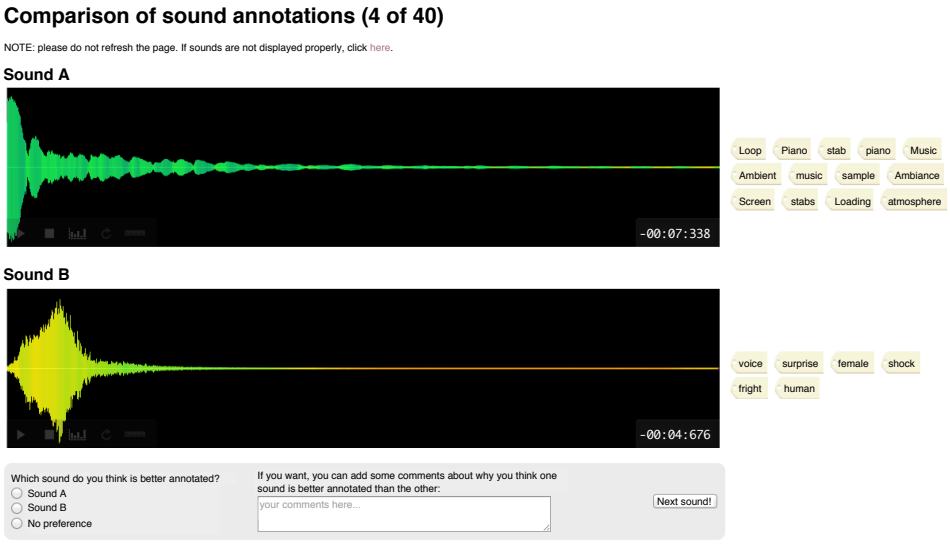

Figure 3: Screenshot of the online experiment interface to judge quality of annotations.

their answers, they also had the option to introduce a textual comment for every comparison.

Participants had to compare the annotation quality of a total of 40 sound pairs. To select the sounds for the experiment, we first randomly chose a set X of 40 sounds among those uploaded after the introduction of tag recommendation. The random selection was only constrained in such a way that all selected sounds had to be uploaded by different users. Then, we built another set Y of 40 sounds uploaded before the introduction of tag recommendation. In order to build Y and make it as similar as possible to X (i.e., containing similar kinds of recordings), we used the "similarity search" functionality of Freesound (Font et al., 2013a). For each sound X_i, we retrieved a list of candidate similar sounds taking into account their acoustic properties represented by low level audio descriptors (note that the similarity search functionality does not take into account any metadata like tags or textual descriptions). Then, we pruned the lists of candidates by removing those sounds that were uploaded after the introduction of tag recommendation and by not allowing to have more than one sound uploaded by the same user. Finally, for each sound X_i, we listened to the remaining candidates and selected the candidate that, in our opinion, was more acoustically similar to X_i. Set Y was thus constructed with all selected candidates. Having the sets X and Y, we formed the final pairs of sounds used in the experiment by iteratively selecting a random sound from each set until we got the 40

pairs determined.

We asked the team of Freesound moderators[3] to participate in the experiment, and collected data from a total of 7 completed experiments (i.e., obtaining a total 7 judgements for every sound pair comparison). Considering the collected data, we assign numerical values to the i-th quality judgement ι_i performed by every participant such that

$$\iota_i = \begin{cases} 1 & \text{if } X_i \text{ is better than } Y_i \\ -1 & \text{if } Y_i \text{ is better than } X_i \\ 0 & \text{if no preference.} \end{cases}$$

Then, qualitative annotation quality ϕ is computed as the average over the union of all quality judgements ι_i performed by all participants in the experiment. Let $\varkappa$ be the union of all quality judgements ι_i, then

$$\phi = \frac{1}{|\varkappa|} \sum_{j \in \varkappa} \varkappa_j.$$

A value of ϕ close to 1 indicates a preference for the annotations of sounds from X (i.e., sounds uploaded after the introduction of tag recommendation), while a value close to -1 indicates a preference for sounds from Y (i.e., sounds uploaded before tag recommendation). A value close to 0 indicates no preference. Our expectation for this metric is to obtain a positive value, indicating a tendency of considering sounds uploaded after tag recommendation as being better annotated than sounds uploaded before tag recommendation. This would suggest an increase in annotation quality.

2.3.3 Cost of the annotation process

- *Average tag application time*: An important indicator of how difficult it is for users to annotate sounds is the observation of the time they spend annotating them (Wang et al., 2012). For that purpose, we define the average time per tag application as

$$\gamma_A = \frac{1}{|A|} \sum_{a \in A} \frac{\lambda_a}{|E^a|},$$

where λ_a is the duration of an annotation session a (in seconds), E^a is the set of tag applications performed during an annotation session a, and A is a set of annotation sessions. Low γ_A values indicate that the

[3]All sounds that are uploaded to Freesound are manually moderated by a small team of people (all of them long-term Freesound users) that ensure the appropriateness of the uploaded sounds. Hence, Freesound moderators are very familiarised with Freesound content and tagging particularities.

time required to add a single tag is lower, therefore it is presumably easier for users to describe sounds.

Unfortunately, the Freesound system did not log information about the duration of annotation sessions before the introduction of tag recommendation, and therefore no data was available for most of the analysis time period. To overcome that issue, during a period of time that lasted two weeks between the March 24 and April 7, 2014, we altered the tag recommendation system so that it only provided recommendations to half of the annotation sessions (but logged the annotation process in both cases). Therefore, our analysis of γ_A is carried out with data gathered during that extra analysis period. This data includes annotation sessions for 562 sounds, one half of them annotated using tag recommendation and the other half annotated without tag recommendation. Note that this new analysis period does not overlap with the period of the main analysis (see below).

We divide the annotation session data we gathered into two sets: one containing data from sessions were tag recommendations were not provided (A^-) and the other containing data from sessions with recommendations (A^+). Next, we compare the average γ for both sets of annotation sessions and asses the statistical significance of the difference by performing the Mann-Whitney U test with a significance level of $p = 0.01$ (Corder and Foreman, 2009). We do not perform any kind of data cleaning or outlier removal over the set of collected annotation sessions. Our expectation for this metric is that sessions which provided tag recommendations will exhibit lower values of γ, as users will add some tags by clicking on the tag suggestions and this will make the annotation process faster.

- *Average percentage of correctly predicted tags*: This metric quantifies how many of the tags assigned to a sound given an annotation process are actually suggested by the recommendation system (thus correctly predicted). Given that the logs of the sound annotation sessions we collected since the introduction of tag recommendation include the lists of all tags that were suggested by the system during the different annotation sessions, we can define the average percentage of correctly predicted tags as

$$\psi_n = \frac{100}{|R_n|} \sum_{r \in R_n} \frac{|T^r \cap T^r_{\mathrm{S}}|}{|T^r|},$$

where T^r is the set of tags assigned to sound r, T^r_{S} is the union of all tags suggested by the system during the annotation process of sound r, and R_n is the total number of sounds uploaded and annotated during the n-th day of our analysis data. Note that we cannot compute ψ for

data before the introduction of tag recommendation. The average percentage of correctly predicted tags is an indicator of the usefulness of the tag recommendation system during the annotation process. High values of ψ indicate that many of the tags that are recommended are actually used to annotate the sounds they are recommended for. Our expectation for this metric is to obtain similar results as in a user-based evaluation of the tag recommendation system we carried out in previous work (Font et al., 2014b). In that case, the average percentage of correctly predicted tags was found to be approximately 33%.

2.4 Analysis methodology

The impact of the tag recommendation system is analysed by looking at the evolution of the Freesound folksonomy (gathering data directly from the Freesound database) and the logs created every time a user annotates a new sound. Our analysis comprises data between September 21, 2011, and February 28, 2014. The tag recommendation system was introduced on November 20, 2013. The metrics defined in the previous section are either computed on a daily basis (using data from a particular day of our analysis), or over bigger periods of time (using data data gathered from several days of our analysis). To represent daily time periods, let $\mathbf{D}$ be a vector of time periods where D_n corresponds to the time period of the n-th day since the beginning of our analysis data. In that vector, D_0 corresponds to the time period of the first day in our analysis data (September 21, 2011), and D_N corresponds to the time period of the last day for which we have analysis data (February 28, 2014). In addition to what precedes, to represent larger periods of time, we define a series of analysis windows which include data from several days of our analysis. On the one hand, let W^I be our analysis window of interest, which represents a time period including all the data after the introduction of tag recommendation (i.e., a total of 100 days from November 20, 2013 to February 28, 2014). On the other hand, let $\mathbf{W}^R$ be a vector of reference analysis windows where each element W^R_m corresponds to a time period of the same length as W^I (100 days), drawn from data before the introduction of tag recommendation. The window W^R_0 corresponds to the last 100 days before the introduction of tag recommendation (from August 12, 2013 to November 19, 2013), and the m-th analysis window corresponds to a time period shifted backwards in time $50m$ days. Figure 4 shows a graphical representation of $\mathbf{D}$ and $\mathbf{W}^R$, and the analysis window of interest W^I. Notice that W^I, as well as each element of $\mathbf{W}^R$, includes a particular range of D time periods (e.g., W^I corresponds to $D_{N-100:N}$).

As mentioned, we are interested in comparing the results of the defined metrics for time periods *before* and *after* the introduction of tag recommendation. In the case of metrics that are computed on a daily basis, we

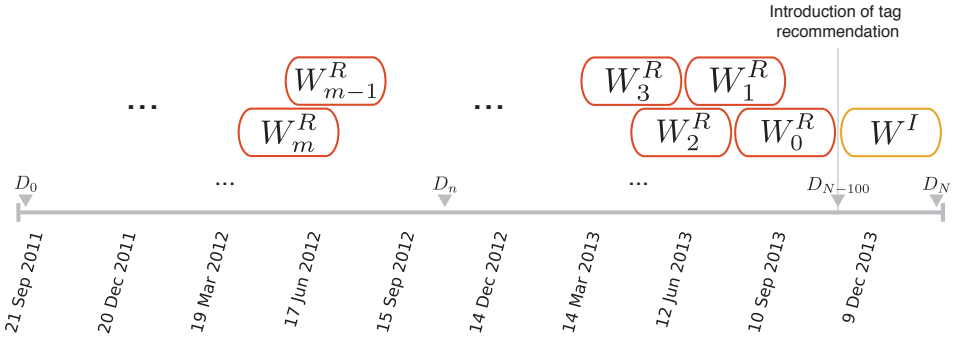

Figure 4: Time period vectors D and W^R, and the analysis window of interest W^I.

perform the comparison by computing the average of each metric over the range of days in **D** included in the window of interest W^I and in each reference window W^R_m. Then, the average obtained from W^I is compared with the average obtained for each time period W^R_m. This results in a total of M comparisons per metric. In our results section, and unless stated otherwise, we always report the results of the comparison between W^I and W^R_m that yields the minimum absolute difference. Hence, our results only show the case in which the tag recommendation system has the least impact. For each one of these comparisons, we assess statistical significance by taking the daily results of the metric corresponding to the compared time periods W^I and W^R_m and performing the Mann-Whitney U test with a significance level of $p = 0.1$. For the case of metrics that are not computed on a daily basis, we follow different approaches for comparing and assessing statistical significance. These approaches have been described for every particular metric in corresponding subsections of Sec. 2.3.

Our analysis data includes annotations for sounds of very different natures and from users with very different levels of expertise. During the analysis period, some users uploaded only one sound, while others uploaded up to 5,500, with the average being on 12.7 uploaded sounds per user. A final point to note is that, although we do not perform any cleaning of the considered Freesound data, we remove from our consideration all tag applications performed by a specific user that, during a narrow time period within W^I (from January 17, 2014 to January 27, 2014), intensively uploaded and annotated sounds using three times more tags per sound than the average. We considered this user as being a clear outlier that could potentially bias the results of our analysis by significantly increasing the average tagline length after the introduction of tag recommendation.

3 Results and discussion

3.1 Vocabulary convergence

3.1.1 Percentage of new tags

Fig. 5 shows the evolution of the percentage of new tags η over the considered time period. We qualitatively see that it decreases after the introduction of tag recommendation. The minimum difference we observe between W^I and all W_m^R is a decrease of 1.7%, which is found to be statistically significant ($p = 4.01 \cdot 10^{-6}$). The maximum difference we observe is a decrease of 5% ($p = 1.26 \cdot 10^{-15}$).

The depicted evolution suggests an influence of the tag recommendation system on the percentage of new tags. However, looking at Fig. 5, a decreasing global trend can be observed, even before the introduction of tag recommendation. To compensate for the existence of such a trend, we perform an extra analysis in which we apply a correction to the η data points obtained from W^I. The correction consists in computing a linear regression with all data points before the introduction of tag recommendation and then subtracting the linear projection of that trend to the data after the introduction of tag recommendation. Once we apply the correction to η over the window W^I, we repeat the comparisons with all reference windows W_m^R and observe, this time, a minimum η decrease of 1.5% which still remains statistically significant ($p = 5.68 \cdot 10^{-5}$).

It could be further argued that during the time period between September 15 and December 14, 2012, a localised decreasing pattern can also be observed with a similar strength to the one we observe after the introduction of tag recommendation. This decreasing pattern might be explained by the apparent local increase that can be observed in the previous months, which might be provoked by a particular user uploading a significant number of sounds with many new tags. Importantly, no relevant patterns can be observed in the other studied metrics during that particular period of time (see below). Moreover, just by simple observation of Fig. 5, it can be spotted that the variance of η is smaller after the introduction of tag recommendation, thus giving more relevance to the observed decreasing pattern in W^I. As mentioned, it is the consideration of similar results from several different metrics that allows us to draw any conclusions regarding the formulated hypotheses.

3.1.2 Average user vocabulary size

Fig. 6 shows the evolution of the average user vocabulary size ς. In it, a clear impact of the tag recommendation system can be observed, as ς consistently increases after the introduction of tag recommendation. When comparing results for the analysis window W^I and the other reference windows W_m^R,

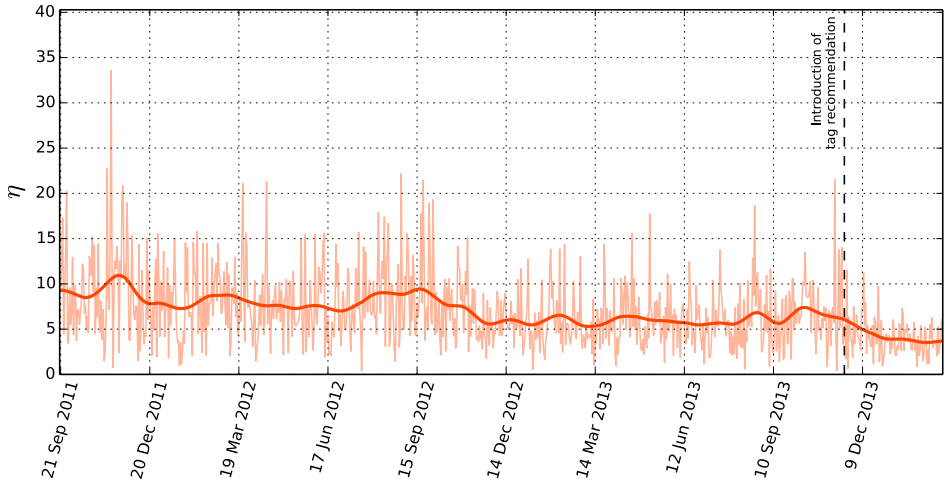

Figure 5: Evolution of the percentage of new tags η. The shaded line corresponds to computed η. The bold line corresponds to a smoothed version of η. Smoothing is performed by convolution over a moving Hann window of 51 days. That particular number of days has been arbitrarily chosen to generate an informative yet visually appealing figure. Unless stated otherwise, the same smoothing strategy is applied in the other figures in this work.

we find a minimum ς increase of 3.46 tags per user ($p = 2.303 \cdot 10^{-11}$). This demonstrates that, after the introduction of tag recommendation, users tend to use a wider variety of tags as their vocabulary size is significantly increased.

3.1.3 User vocabulary sharing

As described in Sec. 2.3, to analyse user vocabulary sharing (μ) we build two networks using data from W_0^R and W^I, respectively. The resulting network built with data from W_0^R has a total of 1,148 nodes (i.e., users) and 73,240 edges, whereas the network built with data from W^I features 1,335 nodes and 122,474 edges. Just by looking at these numbers it can already be seen that users in the W^I network are much more connected. Fig. 7 shows the complementary cumulative node strength distribution of the two networks. The distribution shows that, for a given probability, the network after the introduction of tag recommendation features nodes with a higher strength. Comparing the two distributions yields a statistically significant μ increase of 2.12 ($p = 8.652 \cdot 10^{-17}$). These observations evidence that the tag recommendation system effectively influences users in a way that more tags are shared among them.

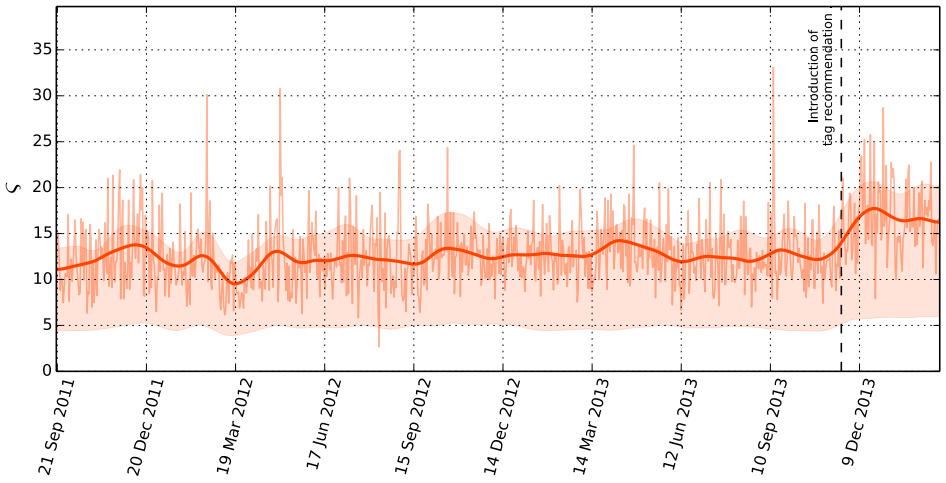

Figure 6: Evolution of average user vocabulary size ς. The shaded line corresponds to computed ς. The bold line corresponds to a smoothed version of ς. The filled area shows the range between the lower and upper quartiles of the original data.

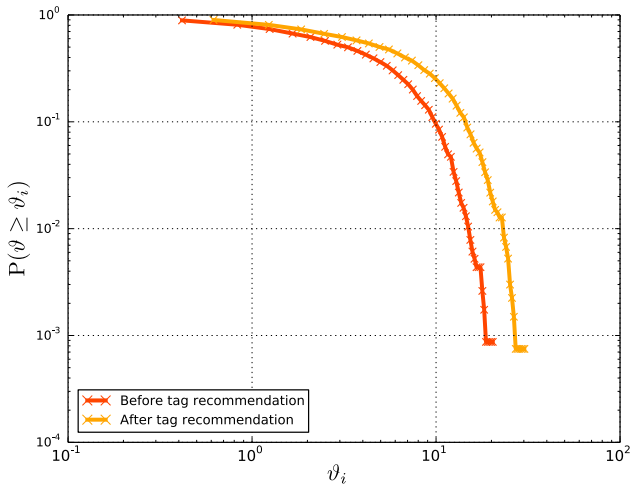

Figure 7: Complementary cumulative node strength (ϑ) distribution of user-user network $\mathcal{U}_k$ before and after the introduction of tag recommendation. Networks are built with data from analysis windows W_0^R and W^I respectively.

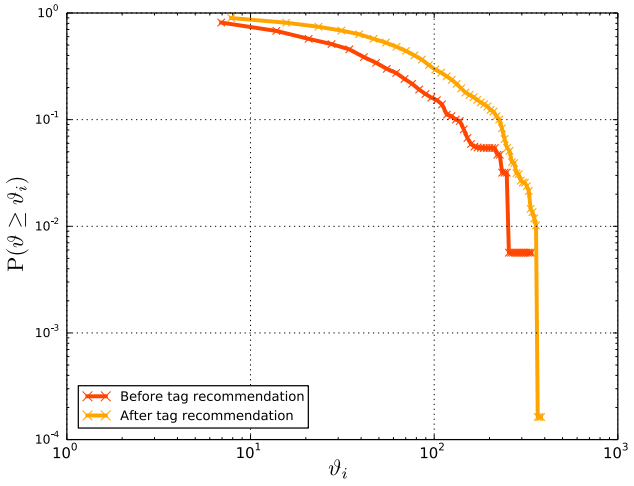

Figure 8: Complementary cumulative node strength (ϑ) distribution of sound-sound network $\mathcal{S}_k$ before and after the introduction of tag recommendation. Networks are built with data from analysis windows W_0^R and W^I respectively.

3.1.4 Sound vocabulary sharing

The analysis of sound vocabulary sharing ν reports similar results to those of user vocabulary sharing. The resulting network built with data from W_0^R has a total of 9,898 nodes (i.e., sounds) and 3,414,449 edges, whereas the network built with data from W^I features 12,946 nodes and 7,405,037 edges. Again, it can already be observed that the network after tag recommendation is much more connected. Fig. 8 shows the complementary cumulative node strength distribution of the two networks. In this case, we also observe an overall increase in node strengths after the introduction of tag recommendation. Interestingly, this is somewhat more relevant in the range of sounds that used to be less connected in the network (roughly for $\nu_x < 200$). The average ν increase is of 34.26 ($p = 2.606 \cdot 10^{-231}$). This result is consistent with what we find in the case of user vocabulary sharing.

3.1.5 Discussion

We have seen that the tag recommendation system diminishes the generation of new tags and, at the same time, it increases the size of users' vocabulary and the number of tags that are shared among users and sounds. This suggests that all users receive a common influence that positively affects the convergence of the vocabulary in the folksonomy by leveraging the reuse of tags, reducing the generation of new ones, and increasing the number of distinct tags in users' personal vocabulary.

We have also found that both user and sound vocabulary sharing are

increased after the introduction of tag recommendation. This observation, combined with the increase in users' vocabulary size, leverages the value of sound annotations. It reveals a better agreement on the vocabulary of tags used to annotate sounds, and also an increase of its size. Therefore, sounds are described using a more coherent and complete vocabulary.

3.2 Quality of annotations

3.2.1 Average tagline length

Fig. 9 shows the evolution of the average tagline length τ. We qualitatively observe a clear increase after the introduction of tag recommendation. Comparing results for the analysis window W^I and reference windows W^R_m, we observe a minimum τ increase of 1.32 tags per sound ($p = 7.553 \cdot 10^{-6}$). Similarly to what we noted in Sec. 3.1.1, Fig. 9 seems to show a global increasing tendency already before the introduction of tag recommendation. We repeated the same extra analysis of that section (i.e., computing the linear regression of data before the introduction of tag recommendation and correcting τ in W^I with the linear projection of the trend) and still observed a statistically significant minimum τ increase of 1.22 tags per sound ($p = 3.65 \cdot 10^{-5}$). Considering the average tagline length for the time periods before and after the introduction of tag recommendation, the observed increase means that sounds are annotated with approximately 20% more tags when users are influenced by the tag recommendation system. This observation is also supported by looking at the histogram of tagline lengths before and after the introduction of tag recommendation (Fig. 10). The increase on the average length of the tagline suggests that annotations using the recommendation system are more comprehensive and, presumably, of better quality than annotations without using the recommendation system.

3.2.2 Percentage of misspelled tag applications

Fig. 11 shows the evolution of misspelled tag applications ω. As expected, we qualitatively observe a slight decreasing tendency in ω after the introduction of tag recommendation When comparing results for the analysis window W^I and the other reference windows W^R_m, we find a minimum ω decrease of 1.4% (not statistically significant), and a maximum decrease of 5% (statistically significant, with $p = 4.775 \cdot 10^{-5}$). Hence, this demonstrates that the introduction of tag recommendation has a moderate impact on misspelled tags, helping users to generate up to 5% less tags with misspellings.

3.2.3 Tag frequency distribution

Fig. 12 shows the complementary cumulative tag frequency distribution before and after the introduction of tag recommendation. It can be observed

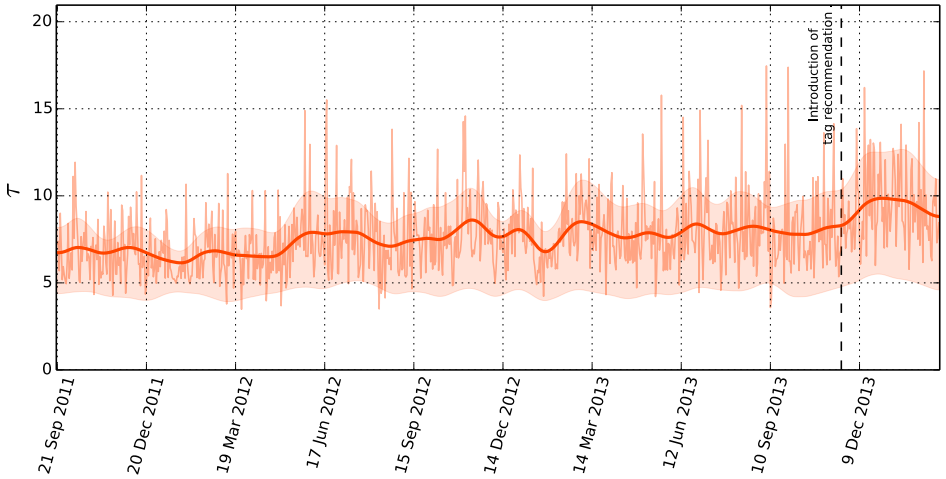

Figure 9: Evolution of average length of tagline τ. Shaded line corresponds to computed τ. The bold line corresponds to a smoothed version of τ. Filled area shows the range between the lower and upper quartiles of the original data.

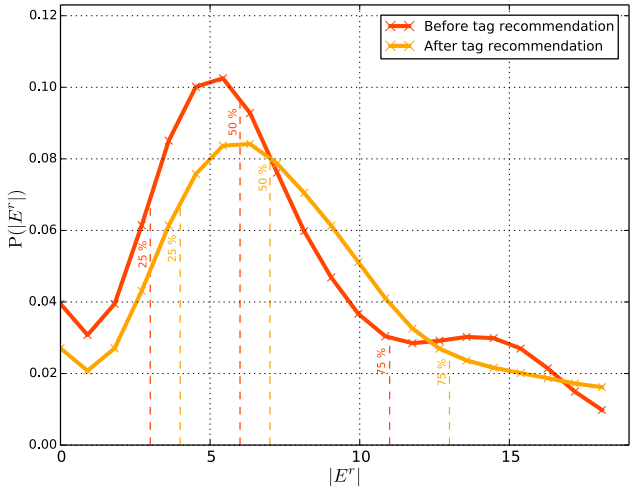

Figure 10: Smoothed normalised histogram of tagline lengths before and after the introduction of tag recommendation. Data is drawn from the analysis windows W_0^R and W^I, respectively. Smoothing is performed using a Hann window of 11 points. Dashed vertical lines with attached percentage values indicate the percentage of sounds whose tagline length is less or equal than that indicated in the corresponding line position.

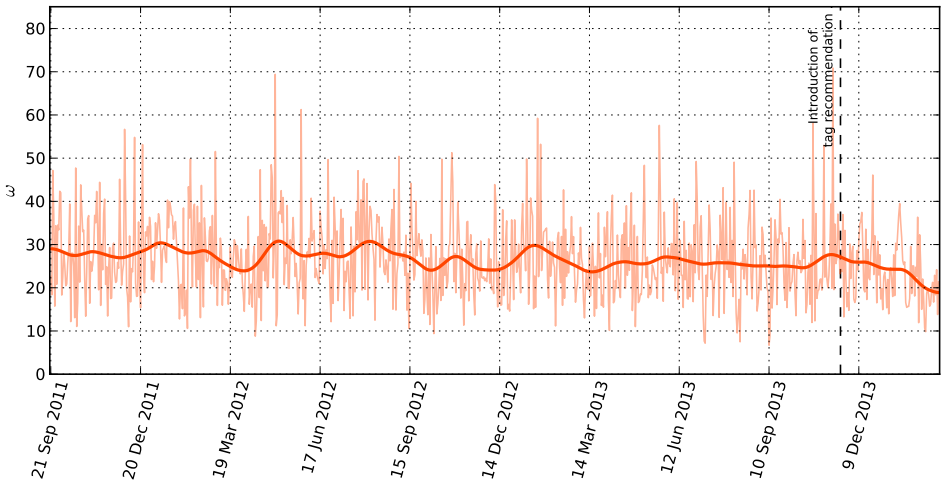

Figure 11: Evolution of the percentage of misspelled tag applications ω. The shaded line corresponds to computed ω. The bold line corresponds to a smoothed version of ω.

that the distribution after the introduction of tag recommendation tends to be more even, particularly reinforcing the usage of tags in the low and mid frequency ranges (tags with less than 800 occurrences). This means that less popular tags gain importance after the introduction of tag recommendation. Less popular tags typically correspond to narrower semantic concepts, which are used to bring more details to sound annotations. Again, this observation is consistent with previous observations regarding vocabulary convergence. It reflects the increase in both user and sound vocabulary sharing, as tags with less frequency gain importance and start being more widely used. It also suggests that annotations after the introduction of tag recommendation are more detailed as usage of tags in the low and mid frequency ranges is reinforced.

To complement these results, we use the method proposed by Clauset et al. 2007 for evaluating how well tag frequency distributions corresponding to the time periods before and after the introduction of tag recommendation fit into a power law distribution[4]. In both cases, the analysis shows that distributions more closely fit a log-normal distribution rather than a power law distribution. However, the tag frequency distribution after the introduction of tag recommendation shows a better fit for the power law than the distribution before tag recommendation, which may also suggest the presence of a better converging vocabulary yielding better quality descriptions (Mathes, 2004; Cattuto, 2006; Halpin et al., 2006; Wagner et al., 2014).

[4]We use an open source implementation as described in Alstott et al. 2014.

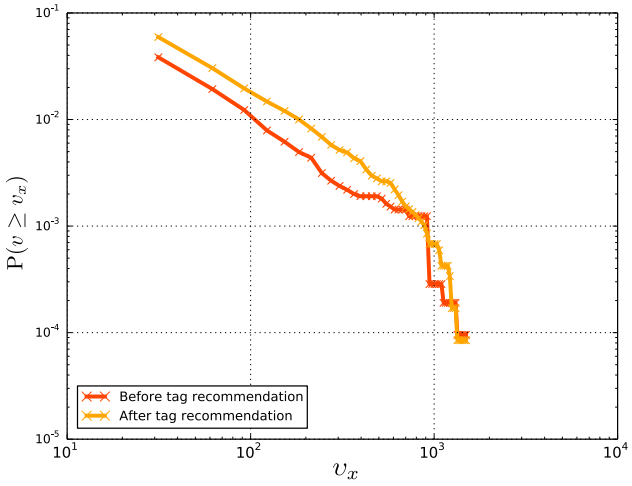

Figure 12: Complementary cumulative tag frequency (v) distribution before and after the introduction of tag recommendation. Data is drawn from the analysis windows W_0^R and W^I, respectively.

3.2.4 Subjective annotation quality

We analyse the results of the online experiment as described in Sec. 2.3.2 and observe a subjective annotation quality $\phi = 0.075$ (0.81 standard deviation). One third of the quality judgements performed by the participants correspond to "no preference" judgements ($\varkappa_j = 0$). If we discard these judgements, the subjective annotation quality is increased to $\phi = 0.114$ (0.99 standard deviation), meaning that in 55% of the judgements, sounds described using the tag recommendation system were considered to be better annotated. These results indicate that participants in the experiment have a slight tendency to consider annotations of sounds described using the tag recommendation system as being better than annotations of sounds made without the tag recommendation system. To further validate these results, we computed Cohen's kappa coefficient to measure the agreement among the quality judgements performed by the participants in the experiment (Carletta, 1996). After pairwise comparisons between the different participants in the experiment, we observe an average kappa coefficient of 0.22. Thus, participants in the experiment tend to agree on their judgements. Overall, this reinforces the previous observations.

Participants of the experiment also provided some textual comments on some of the judgements. In general, participants used comments to explain the reason why they considered sounds to be badly annotated. Among these reasons, the most common ones indicated misleading or uncompleted annotations, the presence of tags not related to the sound being annotated, and the presence of tags with typographical errors. All these reasons are

reported evenly for sounds uploaded before and after the introduction of tag recommendation.

3.2.5 Discussion

We have seen that the average number of tags used to annotate a sound is larger after the introduction of tag recommendation. A similar observation is made in a study by Ames et. al. 2007, in which two mobile phone applications for uploading photos to Flickr[5], an online photo sharing site, are compared. One of the applications features a tag recommendation system to aid users in the tagging process, and an increase in the average tagline length is observed for those photos uploaded with that application.

The fact that the average tagline length increases after the introduction of tag recommendation also reinforces the previously discussed observations regarding the convergence of vocabulary. Tag recommendation yields more tag applications and potentially more comprehensive sound annotations, and yet fewer new tags are created while vocabulary sharing is increased. Hence, our results indicate that sound annotations after the introduction of tag recommendation are done using a more coherent and complete vocabulary of tags. This fact seems to be further confirmed by the results of the online experiment we set up to analyse qualitative annotation quality, as participants on this experiment preferred annotations of sounds uploaded after the introduction of tag recommendation.

The tag frequency distribution we observe after the introduction of tag recommendation also supports the increase in the convergence of the vocabulary. Results indicate that a better agreement is reached specially for those tags with lower frequencies of occurrence. Thus, we could say that there is a better agreement on the tags users choose to annotate specific concepts, which leverages the value (and thus the quality) of the annotations.

Finally, we also observed that tag recommendation helps users in slightly reducing misspellings in the tags they introduce, which also supposes an improvement in the quality of annotations. However, the impact we observe is rather small, which may be explained by several factors. Firstly, the way in which we estimate misspelled tags is not perfectly accurate and thus some noise is present in the metric (Sec. 2.3.2). Secondly, the nature of the tag recommendation system does not prevent itself from actually recommending tags with misspellings. Hence, even if it is intuitively less likely that misspelled tags will feature a strong similarity with any of the input tags, it is still possible that these are recommended. Finally, we can only expect tag recommendation to effectively help in reducing misspellings for the tags that are actually suggested by the system and correctly predicted. As we describe below in Sec. 3.3.2, approximately 19% of the tags of a tagline are

[5] `http://www.flickr.com`

correctly predicted, and this can be taken as a rough estimate of an upper bound for the decrease in the percentage of misspelled tag applications. Furthermore, even when relevant tags are recommended by the system and are correctly predicted, many users still prefer to manually type them instead of clicking on the list of suggestions, which may still lead to misspellings (see Sec. 3.3.2). Overall, our results regarding the quality of annotations suggest that the introduction of tag recommendation has a moderate yet positive impact on this aspect.

3.3 Cost of the annotation process

3.3.1 Average tag application time

Fig. 13 shows the probability density function of the average time per tag application γ with and without the use of the tag recommendation system. Although we observe a smaller average decrease in γ for annotation sessions using the tag recommendation system, it is found to be not statistically significant ($p = 8.3 \cdot 10^{-1}$). This means that there is no substantial difference on the time needed to perform a tag application either using or not using tag recommendation. However, if we look at the total amount of time invested in annotating every sound (instead of every tag), we do observe a statistically significant average increase of roughly 35 seconds per sound after the introduction of tag recommendations ($p = 6.2 \cdot 10^{-3}$), which represents an increase of approximately 20%. This is consistent with the 20% increase of the tagline length we observed in Sec. 3.2.1. In general, we could say that users need at least the same amount of time to perform a single tag application as they needed before using the system. However, annotations are longer and therefore users spend more time annotating sounds.

3.3.2 Average percentage of correctly predicted tags

As explained in Sec. 2.3, the average percentage of correctly predicted tags ψ can only be computed with data drawn from the analysis window W^I. Computing it on a daily basis shows that, on average, approximately 19% (5% standard deviation) of the tags finally assigned to sounds, are suggested by the recommendation system. That observed percentage is 11% lower than the one we found in previous work, where the tag recommendation system was evaluated in a controlled experiment which was not integrated into Freesound (Font et al., 2014b). Hence, we assume this difference is due to the fact that the current analysis is carried out in the real world. Among the correctly predicted tags, we make a distinction between those that are added to the tagline by users clicking on the corresponding tag in the list of suggestions, and those that are manually typed by users. If we only consider the tags that are added to the tagline by actively clicking on the suggestion, we observe an average ψ of approximately 13% (4% standard

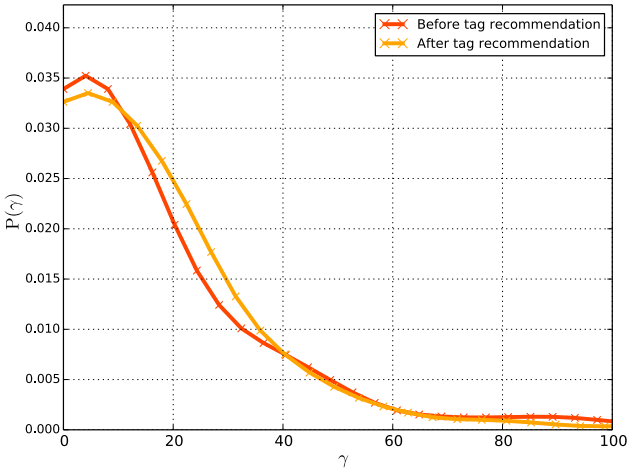

Figure 13: Probability density function of the average time per tag application γ with and without the tag recommendation system. Curves are smoothed using a Hann window of 11 points.

deviation). This means that the tag recommendation system is useful for at least 19% of the annotated tags, but that in many occasions users still prefer to manually type the tags instead of switching to the mouse and clicking on the list of suggestions. In general, these results show that, despite an important part of the final tagline for a sound can be constructed using tags suggested by the recommendation system, the majority of these tags have to be generated by users themselves, and are not necessarily related with those suggested by the system.

3.3.3 Discussion

Contrary to what we expected, we have observed that the tag recommendation system does not seem to have a significant impact on the cost of the annotation process. Although we have seen that users need significantly more time to annotate individual sounds when using the tag recommendation system, we have also seen that this increase can be attributed to the proportional increase of the average tagline length. Hence, the actual time required for every individual tag application does not significantly change. Furthermore, we observed that most of the tags assigned to sounds are not drawn from the list of recommended tags, meaning that most of the annotation process still consists of a generation process where users create tags from scratch rather than a recognition process where users validate tags from a list of suggestions.

There are several potential reasons why we do not observe the expected impact on the cost of the annotation process. On the one hand, we observed that only 13% of the tags in taglines are added from the list of suggestions

by actually clicking on them. Hence, assuming that it is faster to click on tags rather than to manually type them (which is probably not always true), the impact we can expect on the time required for introducing tags should be lower than that 13%. Also, it seems intuitively plausible that users need more time to generate the tags (or recognise them form a list) than to actually introduce them. Hence, the potential impact of lessening the time required for introducing tags is further reduced. On the other hand, the impact of the recommendation system is again limited by the fact that most of the introduced tags are not drawn from system recommendations, and thus an important part of the annotation process does not significantly change after the introduction of tag recommendation. In fact, our results might be suggesting that the cost of the recognition process is not actually lower than the cost of the generation process. This also seems reasonable as the union of all recommended tags for a given sound is much larger than the length of the actual tagline (i.e., new tags are recommended every time that a tag is added to the tagline, see Sec. 2.2), and therefore the recognition process operates over a large set of tags.

Finally, we believe that our metrics regarding the cost of the annotation process are highly dependent on the particular interface of the recommendation system. Also, the recommendation interface can have different impacts according to how users adapt to it. Unfortunately, our analysis does not contain data to be compared coming from other recommendation interfaces. However, to gain some more insight into that aspect, we repeated the calculations of the average tag application time but this time considering experienced and non-experienced users separately. We divided users according to the number of sounds they uploaded during our analysis period. In particular, we set the threshold at the third quartile of the distribution of uploaded sounds per user, which corresponds to 7 uploaded sounds. What we observe is that the average tag application time after the introduction of tag recommendation increases for non-experienced users and decreases for experienced users by a similar amount of about 3 seconds per tag application ($p = 2.15 \cdot 10^{-3}$ and $p = 3.65 \cdot 10^{-3}$ respectively). This shows that experienced users were able to take advantage of the recommendation interface and generate annotations slightly faster, while the interface had a negative impact on non-experienced users, apparently increasing the cost of the annotation process. This could be explained because experienced users probably have a better understanding of the tagging process and can easily interpret and take advantage of tag recommendation. Nevertheless, we think that to draw more consistent conclusions regarding the impact of tag recommendation on the cost of the annotation process, further research should be carried out.

4 Discussion

In this work we have analysed the impact of a state-of-the-art tag recommendation method into the real-world folksonomy of a large-scale system, Freesound. After a conscientious review of current related work, we have identified three main hypotheses regarding the impact that such a method should have when introduced into a collaborative tagging system, and we have defined several metrics to evaluate the impact. We have analysed data comprising of a period from September 21, 2011, to February 28, 2014, the last three months of which correspond to data after the introduction of the tag recommendation method. To the best of our knowledge, these kind of quantitative analyses have not been done before using large-scale data from a real-world folksonomy. Hence, no empirical assessment of the three identified hypotheses was available. The definition of several necessary metrics to assess the three hypotheses is also a further contribution of our work.

Our results show a significant impact of tag recommendation into most of the metrics we defined. However, the result of a single metric in isolation is probably not entirely relevant in our analysis. Instead, the fact that we observe how the changes on several metrics can be explained by some of the outlined hypotheses, gives a particular value to our analysis. Overall, in our scenario, we observe that the first hypothesis (regarding vocabulary convergence) is clearly validated, that the second one (regarding the quality of annotations) only seems to be partially validated, and the third one (regarding the cost of the annotation process) does not seem to be validated. However, we believe the latter is particularly dependent on the annotation interface, and that it could be greatly improved by designing an interface specifically focused on reducing the cost of the annotation process (e.g., favouring clicking on tags rather than typing them), and with a tag recommendation system producing more relevant tag recommendations.

Although in this work we only analyse data in the context of Freesound, we believe that our results are, to some extent, indicative of the impact that tag recommendation can potentially have in other collaborative tagging systems. However, collaborative tagging systems of different nature may react differently to the introduction of a tag recommendation system. An important aspect here is to take into account the motivations that users have for tagging their resources. In narrow folksonomies such as Freesound and Flickr, users typically tag their content so that other users (and also themselves) can easily find it in the future. However, resources are only annotated once, and therefore the tags added by the uploader of a resource must be meaningful to other users of the platform. Contrarily, in *broad* folksonomies such as Delicious and CiteULike[6], resources are tagged multiple times by

[6]Declicious (`http://www.delicious.com`) and CiteUlike (`http://www.citeulike.org`) are two online sharing platforms very popular in the tagging literature, and in which users share bookmarks and scholarly references respectively.

several users, and thus the main motivation for tagging is users' self organisation of the content, without necessarily considering the global context of the sharing platform (Vander Wal, 2005). As a result, very different tagging styles can arise because of the particularities of these two kinds of tagging systems. The tag recommendation system that we use here is designed for narrow folksonomies. It does not try to personalise recommendations to particular users' tagging behaviours, but instead it learns from the whole folksonomy (Font et al., 2014b). Hence, we expect it to have a bigger impact in collaborative tagging systems featuring narrow folksonomies, where the more uniform a tagging style is across users, the better the platform becomes in providing content to other users.

Nevertheless, the metrics and analysis methodology described here are applicable to other collaborative platforms either featuring broad or narrow folksonomies. To further asses the validity of our results, an analysis with data coming from other collaborative tagging systems and tag recommendation systems should be performed. The main obstacle for carrying out this analysis is the limited availability of comprehensive tagging data, including annotations performed *with* and *without* the use of a tag recommendation system, and that comprise user activity for as long a period of time as the one we analysed.

5 Directions for future work

The work presented in this paper points us to several future directions. There are several aspects of the data we already collected that could be further researched to gain more insight into the impact of the tag recommendation system. Firstly, we do not perform any study of the generated taglines at the semantic level. By applying techniques for mapping tags to semantic concepts or categories (e.g., (Cantador et al., 2011)), we could analyse the impact of the recommendation system at the semantic level, and see if it effectively shapes tagging behaviour to a more extensive usage of particular kinds of tags such as content-related or self-organisational tags. Similarly, it could be further researched if other typical problems of tagging systems such as synonymy or polysemy are in fact affected by the use of a recommendation system. Secondly, in the current work we just introduced the concept of user experience when analysing our results in Sec. 3.3.3. It would be interesting to further investigate this aspect by analysing the impact of the recommendation system to other evaluation metrics when considering users with different levels of expertise. Thirdly, another way in which the current study could be further developed would be with the use of network analysis techniques to inspect the user-user and sound-sound networks built on the basis of shared tags. Using such analysis, it would be interesting to evaluate the existence of community structure in those networks and to

see how potential communities in both networks might be related. For example, we could investigate if there are strongly connected communities of users that annotate sounds with a particular tagging style, and then see how the introduction of tag recommendation would affect these communities.

The present work also points out some aspects of tag recommendation systems that should be improved to have a bigger impact on the folksonomies of collaborative tagging systems. In our opinion, the biggest future challenge in tag recommendation is the design of systems that have a bigger impact on the quality of annotations. Annotations are very subjective and difficult to evaluate. However, a recommendation system could be designed to particularly focus on that issue by driving recommendations at higher semantic levels. For example, an intelligent tag recommendation system could analyse the resource being annotated and estimate, on the basis of some domain-knowledge, different information facets that its annotation should cover in order to be "complete". Also, synonymy and polysemy problems could be tackled in tagging systems by suggesting tags to users in combination with alternative variations or disambiguation terms. To produce such recommendations, the recommendation system should probably take advantage of external knowledge bases such as WordNet (Miller, 1995). In order for tag recommendation systems to have a deeper impact in the tagging behaviour and in the quality of annotations in general, we probably need to evolve the basic tag recommendation methods and interfaces to a more complete "assistive" process. In such process, we could better guide users by taking advantage of more knowledge about the semantics of our tags and the particular tagging domain. We foresee that one interesting research direction is the use of ontologies to drive future tag recommendation/assistive tagging systems. Such ontologies should embed knowledge about the domain for which we are recommending tags, including relations between tags and even organising tags into different categories regarding the kind of semantic information they are describing about the resources being annotated.

6 Acknowledgements

The authors would like to thank Perfecto Herrera for his advice during the preparation of this work and also the team of Freesound moderators for participating in the online experiment.

References

Alstott, J., Bullmore, E., and Plenz, D. (2014). Powerlaw: A python package for analysis of heavy-tailed distributions. *PLoS ONE*, 9.

Ames, M. and Naaman, M. (2007). Why we tag: motivations for annotation

in mobile and online media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2007)*, pages 971–980.

Barrat, A., Barthélemy, M., Pastor-Satorras, R., and Vespignani, A. (2004). The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(11):3747–52.

Bischoff, K., Firan, C. S., Nejdl, W., and Paiu, R. (2008). Can All Tags be Used for Search? Categories and Subject Descriptors. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, pages 203–212.

Cantador, I., Konstas, I., and Jose, J. M. (2011). Categorising social tags to improve folksonomy-based recommendations. *Journal of Web Semantics*, 9(1):1–15.

Carletta, J. (1996). Assessing agreement on classification tasks: the kappa statistic. *Computational linguistics*, 22(2):249–254.

Cattuto, C. (2006). Semiotic dynamics in online social communities. *The European Physical Journal C-Particles and Fields*, 37:33–37.

Clauset, A., Shalizi, C. R., and Newman, M. E. J. (2007). Power-law distributions in empirical data. *arXiv:phys*, pages 1–26.

Corder, G. W. and Foreman, D. I. (2009). *Nonparametric Statistics for Non-Statisticians: A Step-by-Step Approach*. Willey.

De Meo, P., Ferrara, E., Abel, F., Aroyo, L., and Houben, G.-J. (2013). Analyzing User Behavior across Social Sharing Environments. *ACM Transactions on Intelligent Systems and Technology*, 5(1).

De Meo, P., Quattrone, G., and Ursino, D. (2009). Exploitation of semantic relationships and hierarchical data structures to support a user in his annotation and browsing activities in folksonomies. *Journal of Information Systems*, 34(6):511–535.

Farooq, U., Kannampallil, T. G., Song, Y., Ganoe, C. H., Carroll, J. M., and Giles, L. (2007). Evaluating Tagging Behavior in Social Bookmarking Systems: Metrics and design heuristics. In *Proceedings of the ACM International Conference on Supporting Group Work*, pages 351–360.

Font, F., Roma, G., and Serra, X. (2013a). Freesound Technical Demo. In *Proceedings of the 21st ACM Conference on Multimedia (ACM MM 13)*, pages 411–412.

Font, F., Serrà, J., and Serra, X. (2013b). Folksonomy-based tag recommendation for collaborative tagging systems. *International Journal on Semantic Web and Information Systems*, 9(2):1–30.

Font, F., Serrà, J., and Serra, X. (2014a). Audio clip classification using social tags and the effect of tag expansion. In *Proceedings of the 53rd AES Conference on Semantic Audio*.

Font, F., Serrà, J., and Serra, X. (2014b). Class-based tag recommendation and user-based evaluation in online audio clip sharing. *Journal on Knowledge Based Systems*, 67:131–142.

Garg, N. and Weber, I. (2008). Personalized, interactive tag recommendation for flickr. In *Proceedings of the 2nd ACM Conference Recommender systems (RecSys 08)*, pages 67–74.

Golder, S. A. and Huberman, B. A. (2006). Usage patterns of collaborative tagging systems. *Journal of Information Science*, 32(2):198–208.

Guy, M. and Tonkin, E. (2006). Folksonomies: Tidying up Tags? *D-Lib Magazine*.

Halpin, H., Robu, V., and Shepard, H. (2006). The dynamics and semantics of collaborative tagging. In *Proceedings of the 1st Semantic Authoring and Annotation Workshop*, pages 1–21.

Ivanov, I., Vajda, P., Goldmann, L., Lee, J.-S., and Ebrahimi, T. (2010). Object-based tag propagation for semi-automatic annotation of images. In *Proceedings of the International Conference on Multimedia Information Retrieval*, pages 497–506.

Jäschke, R., Eisterlehner, F., Hotho, A., and Stumme, G. (2009). Testing and evaluating tag recommenders in a live system. In *Proceedeings of the 3rd ACM Conference on Recommender systems (RecSys 2009)*, pages 369–372.

Jäschke, R., Hotho, A., Mitzlaff, F., and Stumme, G. (2012). Challenges in Tag Recommendations for Collaborative Tagging Systems. In *Recommender Systems for the Social Web*, pages 65–87. Springer Berlin Heidelberg.

Jäschke, R., Marinho, L., Hotho, A., Schmidt-Thieme, L., and Stumme, G. (2007). Tag Recommendations in Folksonomies. In *Proceedings of the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases*, volume 4702 of *Lecture Notes in Computer Science*, pages 506–514.

Li, J. and Wang, J. Z. (2008). Real-time computerized annotation of pictures. In *IEEE transactions on pattern analysis and machine intelligence*, volume 30, pages 985–1002.

Marlow, C., Naaman, M., Boyd, D., and Davis, M. (2006). HT06, Tagging Paper, Taxonomy, Flickr, Academic Article, ToRead. In *Proceedings of the 17th ACM Conference on Hypertext and Hypermedia (ACM Hypertext 2006)*, pages 31–41.

Mathes, A. (2004). Folksonomies Cooperative Classification and Communication Through Shared Metadata. *Computer Mediated Communication - LIS590CMC*, pages 1–13.

Mika, P. (2007). Ontologies are us: A unified model of social networks and semantics. *Web Semantics: Science, Services and Agents on the World Wide Web*, 5(1):5–15.

Miller, G. A. (1995). WordNet: a lexical database for English.

Robu, V., Halpin, H., and Shepherd, H. (2009). Emergence of consensus and shared vocabularies in collaborative tagging systems. *ACM Transactions on the Web*, 3(4).

Sen, S., Lam, S., Rashid, A., and Cosley, D. (2006). Tagging, communities, vocabulary, evolution. In *Proceeings of the 20th Conference on Community Supported Cooperative Work*, pages 181–190.

Sigurbjörnsson, B. and Zwol, R. (2008). Flickr tag recommendation based on collective knowledge. In *Proceedings of the 17th International Conference on World Wide Web (WWW 2008)*, pages 327–336.

Sood, S. C., Owsley, S. H., Hammond, K. J., and Birnbaum, L. (2007). TagAssist: Automatic Tag Suggestion for Blog Posts. In *Proceedings of the 1st International Conference on Weblogs and Social Media (ICWSM 2007)*, pages 1–8.

Spiteri, L. F. (2013). The structure and form of folksonomy tags: The road to the public library catalog. *Information Technology and Libraries*, 26(3):13–25.

Toderici, G., Aradhye, H., Pasca, M., Sbaiz, L., and Yagnik, J. (2010). Finding meaning on youtube: Tag recommendation and category discovery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2010)*, pages 3447–3454.

Turnbull, D., Barrington, L., Torres, D., and Lanckriet, G. (2008). Semantic Annotation and Retrieval of Music and Sound Effects. *IEEE Transactions On Audio Speech And Language Processing*, 16(2):467–476.

Vander Wal, T. (2005). Explaining and showing broad and narrow folksonomies.

Vander Wal, T. (2007). Folksonomy.

Wagner, C., Strohmaier, M., and Huberman, B. (2014). Semantic Stability and Implicit Consensus in Social Tagging Streams. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 735–746.

Wang, M., Ni, B., Hua, X.-S., and Chua, T.-S. (2012). Assistive tagging: A survey of multimedia tagging with human-computer joint exploration. *ACM Computing Surveys*, 44(4):1–24.

Zangerle, E., Gassler, W., and Specht, G. (2011). Using tag recommendations to homogenize folksonomies in microblogging environments. In *Proceedings of the 3rd International Conference on Social Informatics (SocInfo 2011)*, volume 6984 LNCS, pages 113–126.