

RESEARCH

Open Access



Studying tag vocabulary evolution of social tagging systems in learning object repositories

Panagiotis Zervas¹, Demetrios Sampson^{1,2*} and Lina Pelliccione²

* Correspondence: Demetrios.

Sampson@curtin.edu.au

¹Centre for Research and Technology Hellas, Information Technologies Institute, Thessaloniki, Greece

²School of Education, Curtin University, Perth, WA, Australia

Abstract

In the field of Technology-enhanced Learning (TeL), social tagging has been applied to Learning Object Repositories (LORs) mainly as a means:(a) to offer an alternative way of classifying the Learning Objects (LOs) based on the tag vocabulary created by the end-users of the LOs, and (b) to facilitate the enhancement of LOs' descriptions via collaborative tagging. However, in order to be able to understand how a social tagging system performs and whether it can deliver the aforementioned goals, it is important to be able to investigate the evolution of the tag vocabulary, which constitutes the core component of a social tagging system. Within this context, research has focused on different facets of social tagging systems such as the growth of the tag vocabulary, the frequency and reuse of tags, as well as the stability of the tag vocabulary but there are only sporadic studies for investigating these issues in the field of LORs. This paper aims to contribute in studying how social tagging systems perform in the context of LORs by investigating the evolution of the tag vocabulary in OpenScienceResources Repository, a science education domain specific repository with a rich dataset operating in Europe for 5 years.

Keywords: Social tagging system, Learning object repository, Tag vocabulary, Tag growth, Tag reuse, Tag discrimination

Introduction

The emerging Web 2.0 applications have allowed for alternative ways of characterizing digital resources, which move from the expert-based descriptions following formal classification systems to a more informal user-based tagging (Hsu et al. 2014; Derntl et al., 2011; Bi et al., 2009). This alternative way of characterizing digital resources is referred to as “*social tagging*” and it is defined as the process of adding keywords, also known as tags, to any type of digital resource by the users rather than the creators of the resources (Hammond et al., 2005; Heymann et al., 2008). The collection of tags which is created by the different users is referred to as “*tag vocabulary*” (Smith, 2008; Golder & Huberman, 2006). Even though user-generated tags pose specific limitations, including synonymy, ambiguity and typographical errors (Ma, 2012), social tagging has been extensively explored due to its potential to enhance traditional classification methods of digital resources in the web. More precisely, it has been argued that social tagging can facilitate the generation of massive amount of tags reflecting “the wisdom

of crowds". As a result, it is anticipated that the generated tag vocabulary could be a promising and more relevant to the users supplement (superset or subset) of the corresponding existing taxonomies adopted by the metadata experts (Ma, 2012).

Social tagging has also received attention in the field of Technology-enhanced Learning (TeL) mainly due to the emergence of Open Educational Resources (OERs) initiatives worldwide, which have focused on supporting the process of organizing, classifying and storing digital educational resources in the form of Learning Objects (LOs) and their educational metadata in web-based repositories referred to as Learning Object Repositories (LORs) (Ehlers, 2011). Furthermore, social tagging in TeL has recently been also considered for other purposes, such as supporting student assessment (e.g., Kardan et al., 2016) or supporting the provision of personalized learning objects and pathways to students (e.g., Cao et al., 2015). However, these approaches are still not widely adopted.

In a recent study of 49 well-known LORs (Zervas et al., 2014), it was reported that 27 % of them are using social tagging systems, so as to enable their end users (namely teachers and students) to characterize the LOs hosted in these LORs with their personal tags. Applying social tagging in LORs could offer the following benefits (Zervas & Sampson, 2014): (a) an alternative way for classifying and navigating to LOs based on tag vocabularies generated by end-users and not only by an externally defined classification system, (b) a mechanism to facilitate the enhancement of LOs descriptions via collaborative tagging, so that eventually LOs will not only carry their creators' anticipated contextual value but also different end-users' contextual value.

Both these benefits of social tagging when applied to LORs aim to enrich the LO descriptions with information potentially useful to teachers either in terms of the content of the LO (e.g., subject domain concepts described by the LO) or in terms of how the LO can be used in teaching and learning (e.g., teachers' experiences from using the LO in their teaching practice). In this way, teachers within an online community can be facilitated to search, identify and select LOs that are not only meaningful to them based on their content, but also in terms of relating to their own teaching needs and context.

Within this context, in order to be able to understand how a social tagging system performs and whether it can deliver the aforementioned benefits to its users, it is important to investigate the evolution of the tag vocabulary, which constitutes the core component of a social tagging system (Ma, 2012). Many studies have been conducted on different aspects of social tagging systems such as the growth of the tag vocabulary, the frequency and reuse of tags, as well as the stability of the tag vocabulary (Santos-Neto et al, 2013; Ma, 2012; Robu et al., 2009; Farooq et al., 2007; Golder & Huberman, 2006), but the vast majority of studies utilize only the tag vocabulary growth metric, neglecting other metrics. Furthermore, in the context of TeL there are only sporadic studies for investigating these issues in the field of LORs, also mainly focusing on the tag vocabulary growth metric. However, as aforementioned, social tagging in LOR aims not only to provide the means for better organizing and classifying LO, but also a means for teachers to infuse their actual experiences in the LO description and better support search and retrieval for their peers, from this perspective. Therefore, further works should be conducted to understand the behavior of social tagging systems within LORs, and more specifically, focusing on different types of learning objects accommodated in these LORs. Furthermore, additional metrics, such as tag re-use and tag discrimination should be included in these works, since they can

offer deeper insights to the behavior of the social tagging system, complementing the tag vocabulary growth metric.

In this context, this paper aims to contribute in the under-researched aspect of studying how social tagging systems perform in the context of LORs by (a) investigating the case of the OpenScienceResources Repository, a science education domain specific repository comprising diverse types of LOs, with a rich dataset operating in Europe for 5 years and (b) adopting a wide range of metrics to study the behavior of the social tagging system and the evolution of the tag vocabulary, namely tag vocabulary growth, tag re-use, tag discrimination and tag entropy. This can lead to more informed design considerations for the incorporation of social tagging features in large-scale repositories of educational resources.

Following this introduction, the rest of the paper is organized as follows. Background discusses the concept of social tagging, its expected benefits and provides an overview of related studies that investigate the dynamics of social tagging systems with an emphasis on the analysis of the evolution of the tag vocabulary, both within LORs as well as in repositories outside TeL. In Research method, we present the research method used in our study, in which the data collection process from an existing LOR, namely the OSR Repository and the research methodology with specific metrics are introduced for investigating the evolution of the tag vocabulary. In Results and discussion, we present the results from the application of our research methodology and we discuss our findings. Finally, the paper concludes with the practical implications of the results, as well as potential future research directions in this field.

Background

Social tagging of learning objects

Learning Objects (LOs) are a common format for developing and sharing educational content and they have been defined by Wiley (2002) as: “*any type of digital resource that can be reused to support learning*”. LOs and their associated metadata are typically organized, classified and stored in web-based repositories which are referred to as Learning Object Repositories (LORs) (McGreal, 2008). The majority of the LORs that are currently operating online adopt the IEEE LOM standard (IEEE LTSC, 2005) or an application profile of IEEE LOM (Smith et al. 2006) for describing their LOs aiming to facilitate search and retrieval of them among different LORs (McGreal, 2008).

Despite the use of well-defined formal metadata for LOs, the users of the LOs (that is, teachers and students) are facing difficulties to discover and find suitable LOs from LORs (Hyon, 2011; Dahl & Vossen, 2008; Al-Khalifa & Davis, 2007). This has led to the investigation of other means for describing LOs such as social tagging (Bateman et al., 2007; Hammond et al., 2005). With social tagging the creators of metadata need no longer to be metadata experts or the authors of the LOs. Instead, the generation of metadata is done by the end-users of the LOs, who can describe educational resources with tags that are meaningful to them and that can facilitate users’ searching and retrieval of previously used and already known LOs (Doush, 2011; Huang et al. 2011). The expected benefits of this approach can be summarized as follows:

- LOs are labeled with users’ personal tags, which reflect their personal way of describing, classifying, locating and navigating to LOs. This could offer a

personalized way for searching which is delivered by users' tags and not by an externally defined classification system (Cho et al., 2011; Vuorikari et al., 2010)

- LOs are tagged by different users with an increased amount of tags that reflect "the wisdom of crowds". This could offer a mechanism to capture users' contextual value of LOs (e.g., experiences from using the LO in teaching and learning practice), which could be different from creators' anticipated contextual value (Zervas & Sampson, 2014; Trant, 2009; Dahl & Vossen, 2008).

However, in order to be able to understand how a social tagging system performs and whether it can deliver the aforementioned benefits to its users, it is important to investigate the evolution of the core component of a social tagging system, namely the tag vocabulary (Ma, 2012). Next, we discuss existing works that are relevant to the scope of our study and mainly focus on analyzing and studying the behavior of social tagging systems and the evolution of the tag vocabulary.

Related studies: analysis of the Tag vocabulary of social tagging systems

Several studies have been undertaken to study the evolution of social tagging systems' tag vocabulary. Early research conducted by Golder & Huberman (2006), who investigated the tagging dynamics of del.icio.us (2016). More specifically, the authors studied the growth of the tag vocabularies of specific users and they showed that these vocabularies are continually growing and evolving over time. Moreover, the authors demonstrated that this continuous growth of the tag vocabularies of specific users is related to the discovery by these users of new items (here: bookmarks) and the addition of new tags to categorize and describe them. Marlow et al. (2006) have studied the growth of tag vocabulary over time for the case of Flickr (2016). More specifically, the authors showed that the addition of new tags is strongly correlated with the addition of new items (here: photos) and it is also only moderately correlated with the registration of new users to the system. Cattuto et al. (2007) analyzed the growth of the global tag vocabulary (i.e. the cardinality of the set of distinct tags within the social tagging system) and the growth of local tag vocabularies (namely the growth of distinct tags addressed at a specific resource or generated by a given user) of del.icio.us. The authors reported that the growth trend followed a power law distribution (exponent of 0.8) at the global level and sub-linear growth of the local tag vocabularies for specific resources and users. This difference has been explained by the authors to be related with different users' tagging behavior. In another study, Farooq et al. (2007) studied social tagging dynamics of CiteULike (2016) and proposed six tag metrics, namely growth, reuse, non-obviousness, discrimination, frequency, and patterns, so as to explain the dynamics of the CiteULike system. The authors measured the cumulative number of new tags generated each month and they concluded that new tags are perfectly correlated with the new users registered to the system. They demonstrated also that most of the tags were generated by a relatively small group of users and a significant set of tags was not reused, whereas few tags were reused a significant number of times. Chi & Mytkowicz (2008) analyzed the social tagging activities of del.icio.us and they proposed a metric based on information entropy for drawing insights about the tagging behavior of del.icio.us users. More specifically, the authors calculated the entropy of tags, the entropy of

documents, and the entropy of users, as well as the entropy of documents conditional on tags and the entropy of tags conditional on documents. Based on their results, they concluded that over time the users were heavily reusing eachothers' tags and thus, the navigation afforded by the social tags in the system was reduced. Robu et al. (2009) studied the tag distributions from 500 websites collected from del.icio.us and examined the top 25 tags for each. The authors reported that the websites that contained a larger number of tags followed a power law distribution. Makani & Spiteri (2010) selected three metrics proposed by Farooq et al. (2007), namely tag growth, tag reuse, and tag discrimination, to examine the evolution of the tag vocabulary of the knowledge management community of interest in CiteULike. Their results indicated a steady decrease in the number of unique tags over time, suggesting an increasing stability in the community vocabulary and the establishment of domain-specific vocabulary. Moreover, community members highly reused eachothers' tags over time and demonstrating increased collaboration in this matter. In another study, Ma (2012) focused their research on identifying the factors which affected the growth of distinct tags of a given resource within the context of CiteULike. Furthermore, the authors also investigated how this growth progresses overtime and whether it reaches a point of stability. The author reported that the ratio of the distinct tags for a given article over the total tags is highly dependent from three factors, namely the cardinality of the user set who have assigned a tag to the article, the date that the article was initially tagged and the life span of the article. Finally, Santos-Neto et al (2013) studied whether growth of users' tag vocabularies changes according to the user age. The study was conducted with data from three different social tagging systems, namely CiteULike Connotea and del.icio.us. The results indicated that users' tag vocabularies are constantly growing, but at different rates depending on the age of the user.

In summary the previous studies showed that: (a) the tag vocabulary is growing over time (following power law distributions) until a stabilization point, which indicates the maturity of the vocabulary within the users' community of the social tagging system, (b) the growth of the tag vocabulary could be affected by several factors such as the number of new resources entered in the system, the number of new users registered to the system, the users' age in the tagging system, as well as the life span of the resources in the tagging system and (c) the further analysis of the tag vocabulary with appropriately selected metrics can provide insights about the tagging behavior of a social tagging system's users. Our work complements and extends these studies as it investigates the dynamics of social tagging systems applied in LORs. Moreover, the application field of LORs provides a unique opportunity to investigate whether the evolution of tag vocabulary is affected by the different educational resources (LO) types hosted in LORs, namely images, videos, references and readings, simulations, as well as teachers' guides and lesson plans. This is important since a prevailing aspect among current studies is that they perform analysis of tag vocabularies applied to a specific type of resources (such as websites in case of del.icio.us, academic papers in case of CiteULike, photos in case of Flickr).

Within the TeL literature, there are limited and sporadic studies, which have investigated the dynamics of social tagging systems applied in LORs. A relevant study has been performed by Vuorikari & Ochoa (2009), who investigated the distribution of tags per month, the tag growth and the tag reuse of the Calibrate Portal¹ (2016). The results demonstrated that tag growth is strongly correlated with the registration of new users

to the portal. Moreover, tag reuse was very low and the authors reported that this might have been influenced by the tagging interface where popular tags were absent. Nevertheless, the authors do not consider other metrics in their study (such as tag discrimination) for further analyzing the tag vocabulary of the Calibrate Portal towards gaining insights about the tagging behavior of Calibrate Portal's users. Additionally, this study does not consider aspects of the tag vocabulary growth in relation to the different LO types included in the Calibrate Portal. Our study complements and extends the study of Vuorikari & Ochoa (2009) by: (a) applying additional tag metrics for analyzing the evolution of the tag vocabulary in an interrelated manner towards drawing insights about the tagging behavior of the users of a social tagging system applied in a specific LOR and (b) investigating the evolution of the tag vocabulary for different LO types.

Research method

Data collection and normalization

This research is based on data produced in OpenScienceResources (OSR) Repository (2016) for over 4,5 years, namely from 1 November 2009 until 31 May 2014. The OSR Repository was developed in the framework of an EU-funded project, referred to as "OpenScienceResources: Towards the development of a Shared Digital Repository for Formal and Informal Science Education" (2016). It provides access to openly licensed (through Creative Commons) science education LOs, which can be exploited by science teachers in their day-to-day science teaching activities, connecting formal science education in schools with informal science education activities taken place in European Science Centres and Museums (Sampson et al., 2011b).

The science education LOs that are included in the OSR Repository have been characterized: (a) with educational metadata by the content providers (namely, European Science Centres and Museums, as well as science teachers) following an application profile of the IEEE LOM standard (Sampson et al. 2011a, 2011b) and (b) with social tags by the end-users of the repository (i.e. science teachers) and with the use of a social tagging tool, namely the ASK Learning Objects Social Tagging (ASK LOST 2.0) (Sampson et al., 2011a). It is worth mentioning that registered users of the OSR Repository comprises mainly of science teachers, who are able to upload and share their LOs with other users and/or search and find appropriate LOs for their day-to-day science teaching activities. ASK LOST 2.0 is a web-based tool for socially tagging LOs, which has been integrated to the OSR Repository. The main functionalities of the ASK LOST 2.0 include (Sampson et al., 2011a):

- LOs tagging: The user can characterize with his/her selected tags any kind (URL or digital file) of science education LO. The tags that the user can add to the science education LOs describe the topic and/or the subject domain of a science education LO related with the science curriculum.
- Guided Tagging: During the tagging process of a science education LO, the user is presented first with his/her tags previously used for characterizing other science education LOs (referred to as personal tags) and then with tags that are most frequently used by other users regarding this specific science education LO (referred to as popular tags).

- Auto-Suggested Tagging: During the tagging process, the user is presented with suggested tags that have been used by other users and are relevant with the tag that the user is typing.
- Creation of user's personal LOs collection: The user has the capability to save to his/her personal list, science education LOs uploaded by other users and browse the tags that these users have used.
- Browse LOs via tag cloud: The user can search and browse science education LOs using an appropriately formatted tag cloud produced by the tags that all users of the tool have offered. The tags that have been previously used by the user are presented with red color within the tag cloud.

In order to address the issue of reliability and validity of the social tags that were analyzed in our study, we applied the following data cleaning methods as they have been proposed by Golder & Huberman (2006): (a) we corrected tags with grammatical errors, (b) we removed tags that were irrelevant with the content of the LOs, such as tags used to express end-users' opinions and/or emotions like funny, cool, amusing, etc., (c) we removed tags that were synonyms with other tags and (d) we translated to English tags that had been added in other languages. This also means that if a tagger had only contributed tags that were irrelevant with the content of the LOs or tags that were synonyms with other tags then this tagger was excluded from our study.

Table 1 presents the snapshot of the cleaned OSR Repository dataset for the data from 1 November 2009 to 31 May 2014.

As we can notice from Table 1, during our study the OSR Repository included 11.175 social tags (2.735 of them were distinct), which had been added to 2.018 science education resources. This means that, on average, approximately 5 social tags were added per science education LO (1 of them is distinct and 4 of them are duplicates).

Methodology

In our research methodology, we have adopted three main tag metrics that have been proposed by Farooq et al. (2007). Further to that, we propose how these main tag metrics could be interpreted and combined with other metrics, so as to be able to provide meaningful insights about the evolution of the tag vocabulary of a LO social tagging system. Next, we present the tag metrics used in our research methodology:

Tag growth

This metric aims to visualize the creation of new tags over time. Analyzing tag growth in a social tagging system provides an index of how the tag vocabulary is evolving over time. This metric could provide answers to questions about the rate of creation of new tags, as well as whether the vocabulary is stabilizing over time. However, in order to be

Table 1 OSR Repository Dataset (1 November 2009 to 31 May 2014)

| Variables | Value |
|------------------------------------|-------|
| Taggers | 321 |
| Tagged Science Education Resources | 2018 |
| Social Tags | 11175 |
| Distinct Social Tags | 2735 |

able to identify whether the creation of new tags is influenced by other factors, we need to combine this metric with other metrics, namely the rate of new users registering to the OSR Repository, as well as the rate of new LOs added to the OSR Repository. Additionally, in order to be able to fully understand the dynamics of a social tagging system, the tag growth metric should be combined with the *entropy of tags*. Entropy measures the amount of uncertainty about a particular event associated with a probability distribution (Shannon, 2001). Thus, entropy of tags depicts how much new information each tag contains (compared to the rest of the tag set), therefore making the associated LO easier to search and retrieve since it is assigned with a 'rare' tag. Entropy of tags can be calculated using the following formula as proposed by Chi & Mytkowicz (2008):

$$\mathbf{H}(\mathbf{Tag}) = \sum_{i=1}^N p_i \log(p_i) \quad (1)$$

where $p(i)$ is the probability of the i th tag of the tag vocabulary to occur within the set of total tags and N is the number of tags of the tag vocabulary. Based on the above formula, there are two main cases in which entropy of tags can change: (a) the total number of tags of the tag vocabulary increases then the entropy will increase and (b) the tag probability distribution becomes more uniform then the entropy will also increase. In the former case, this means that users are adding distinct tags to the LOs of the repository, whereas in the latter case the users are reusing tags that are relatively not popular in the social tagging system. As a result, by combining tag growth and the entropy of tags, we can extract conclusions about the behavior of a social tagging system.

Finally, it should be mentioned that LOs could be of different types such as videos, simulations, images etc., as defined by the OSR LOM application profile (Sampson et al., 2011b). Based on that, we can calculate the tag vocabulary growth rate for each LO type. This will enable us to identify whether specific LO types can achieve higher tag vocabulary growth rates than other types. Tag growth rate, which depicts the rate in which tag vocabulary for each LO type is evolving, can be calculated using the following formula as proposed by Strohmaier et al (2012):

$$\mathbf{TagGrowth\ Rate} = \frac{\# \text{ of Tags for specific LO type}}{\# \text{ of LOs of specific type}} \quad (2)$$

Tag reuse

this metric examines the level to which existing tags are being used by users to characterize LOs instead of creating new tags. Tag reuse can be calculated using the following formula as proposed by Farooq et al. (2007):

$$\mathbf{TagReuse} = \frac{\sum \# \text{ of Distinct Users for each Tag}}{\# \text{ of Tags}} \quad (3)$$

Considering that each tag will have at least one associated user, the minimum value of tag reuse is 1.0 user/tag. Tag reuse provides a direct interpretation of how often tags in a social tagging system are being recycled among the users. Both tag growth and tag reuse are important metrics to understand how the tagging vocabulary evolves and how the social tagging system behaves. More specifically, a social tagging system could have:

- *High tag growth and low tag reuse*: this means that users are mainly adding new tags and they are not re-using existing tags. As a result, the specificity of tags is increasing and this could facilitate users to narrow their search results when using specific tags.
- *Low tag growth and high tag reuse*: this means that the social tagging system is highly collaborative and LOs' tags are increased over time. However, the specificity of tags is decreasing and any single tag references many LOs. In this case, average number of tags used in a search query should be increased by the users in order to narrow the search results.
- *High tag growth and high tag reuse*: this means that the users are both adding new tags and re-using existing tags. In this case, tag growth and tag reuse should be examined in combination with other metrics (such as tag discrimination that is described below), so as to be able to interpret the behavior of the social tagging system.
- *Low tag growth and low tag reuse*: this means that for some reason the system is not used at all for tagging by its users.

Additionally, tag reuse can be calculated for different LO types. This will facilitate us to combine this metric with tag growth metric and extract conclusions about the behavior of each LO type within the social tagging system of the OSR Repository.

Finally, in order to be able to compare the social tagging system of the OSR repository with other social tagging systems, we need to plot the distribution of tags' reuse occurrences per number of tags, as well as the distribution of tags reuse occurrences per number of users. Previous studies have observed that these distributions resemble a power law (Robu et al. 2009; Cattuto et al., 2007; Farooq et al. 2007) and it will be interesting to demonstrate similarities with these studies.

Tag discrimination

the aim of this metric is to calculate for individual tags their discriminating value, namely how well they discriminate the resources that have assigned to. Tag discrimination value can be calculated by considering the ratio of the number of distinct LOs which have been assigned each tag to the number of all tags. This depicted in the following formula as proposed by Farooq et al. (2007):

$$\text{TagDiscrimination} = \frac{\sum \# \text{ of Distinct LOs for each Tag}}{\# \text{ of Tags}}$$

The tag discrimination metric can be helpful if monitored over time, so as to provide insights about the usefulness of tags over time in their ability to discriminate among LOs of a LOR. Tag discrimination can also be calculated for the different LO types. This could facilitate us to identify whether the LO type affects the discriminative value of tags.

Results and discussion

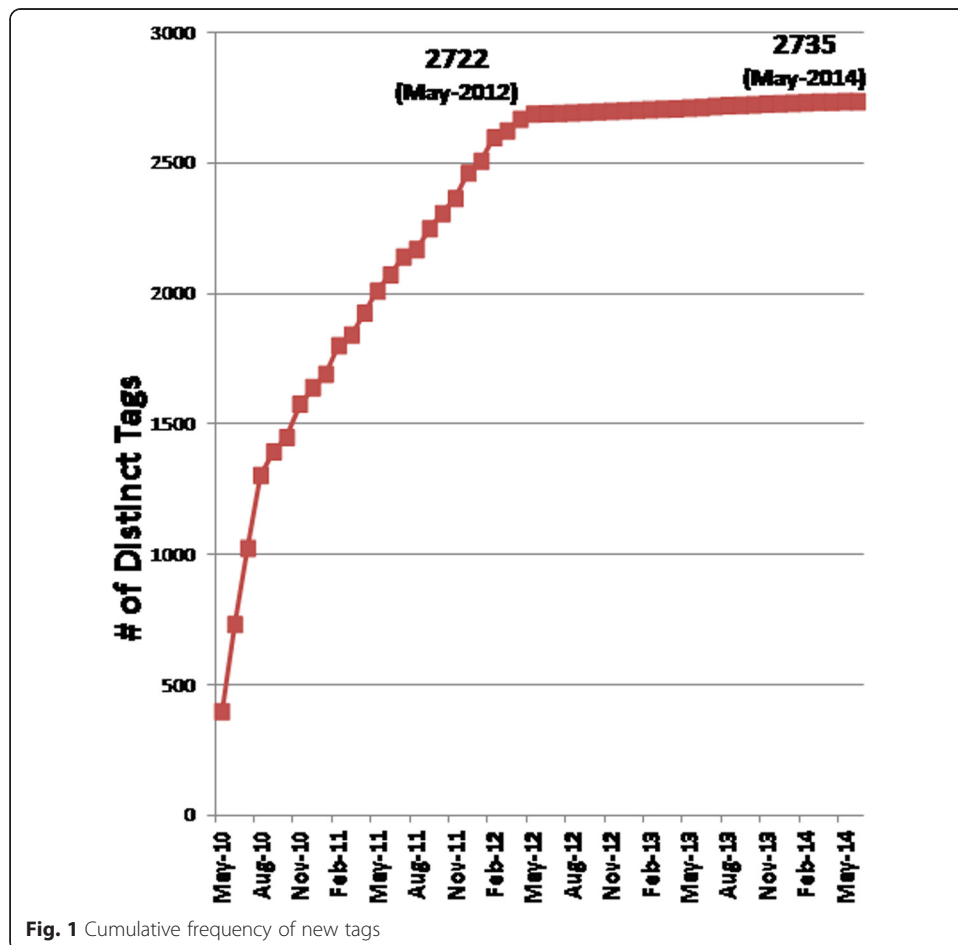
Analysis of tag growth

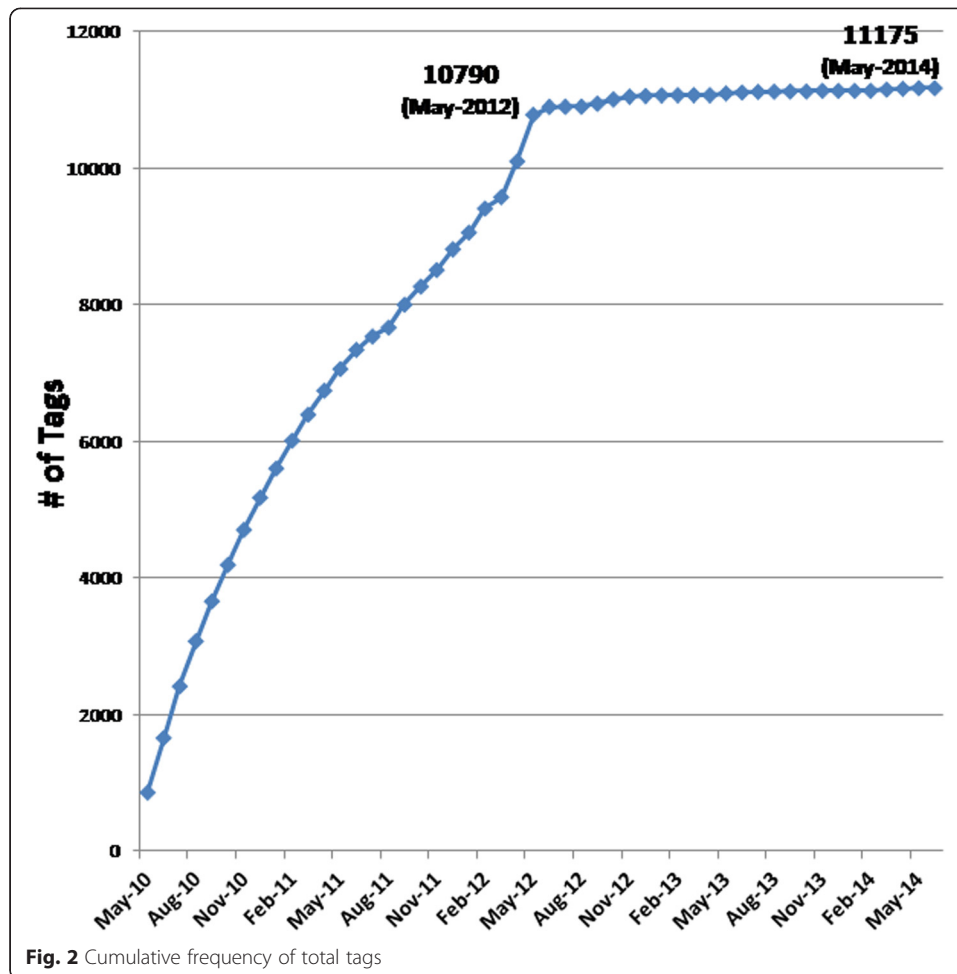
In order to analyze the tag growth of the OSR repository, we calculated the number of distinct tags and the number of total tags created per month starting from May 2010 (namely the month that the first tags from OSR users were created) until May 2014

(namely the last month included in our dataset). Figures 1 and 2 present the cumulative frequency of new tags and total tags in the OSR Repository over time correspondingly.

As we can notice from Fig. 1, there is a high tag growth until May-2012. After that time, it seems that the tag vocabulary is stabilized (namely, no new tags are added to the OSR Repository), although the total tags slightly keep increasing (as depicted in Fig. 2). In order to be able to identify which factors affected this stabilization point, we calculated (a) the number of new users registered to the OSR Repository starting from November 2009 (namely, the first month that users were registered to the repository) until May 2014 (namely, the last month included in our dataset) and (b) the new LOs added to the OSR Repository starting from January 2010 (namely the first month that LOs were added to the repository) until May 2014 (namely the last month included in our dataset). Figures 3 and 4 present the cumulative frequency of new users and LOs over time correspondingly.

From Fig. 3, we can notice that there is a high increase of new users registering to the system (OSR repository) until May-2013 and after that date it appears that only a limited number of new users are registering to the system. As Fig. 4 depicts, new LOs are also being added at a high rate until May-2012 and after that date it seems that only a limited number of new LOs are being added to the repository. Based on these data, we can deduce that the reason for the stabilization of the tag vocabulary on May-2012





could be related to the relative low number of new LOs and/or users being added to the system (OSR Repository) after that date.

To further support and verify this assumption, we performed a correlation analysis (using the Pearson's correlation coefficient) between (a) the number of new tags added per month and the number of new users registered per month, as well as (b) the number of new tags added per month and the number of new LOs added per month. Table 2 presents the results of the correlation analysis, namely the calculated Pearson correlation coefficient r .

As we can notice from Table 1, there is a statistically significant strong correlation ($r = 0,545$, $p < 0,001$) between the number of new tags added in the system and the number of new LOs uploaded in the system. These results validated our initial assumption that new tags are strongly influenced by the addition of new LOs to the OSR Repository. Furthermore, there is a statistically significant weak correlation ($r = 0,287$, $p < 0,05$) between the number of new tags added in the system and the new users added to the system. This means that the number of new users being registered to the system influenced the addition of new tags, however the impact of this influence was weaker than the impact of new LOs. A possible reason for this is that the OSR Repository is a science education domain-specific repository and its users are European school science teachers. This means that, the

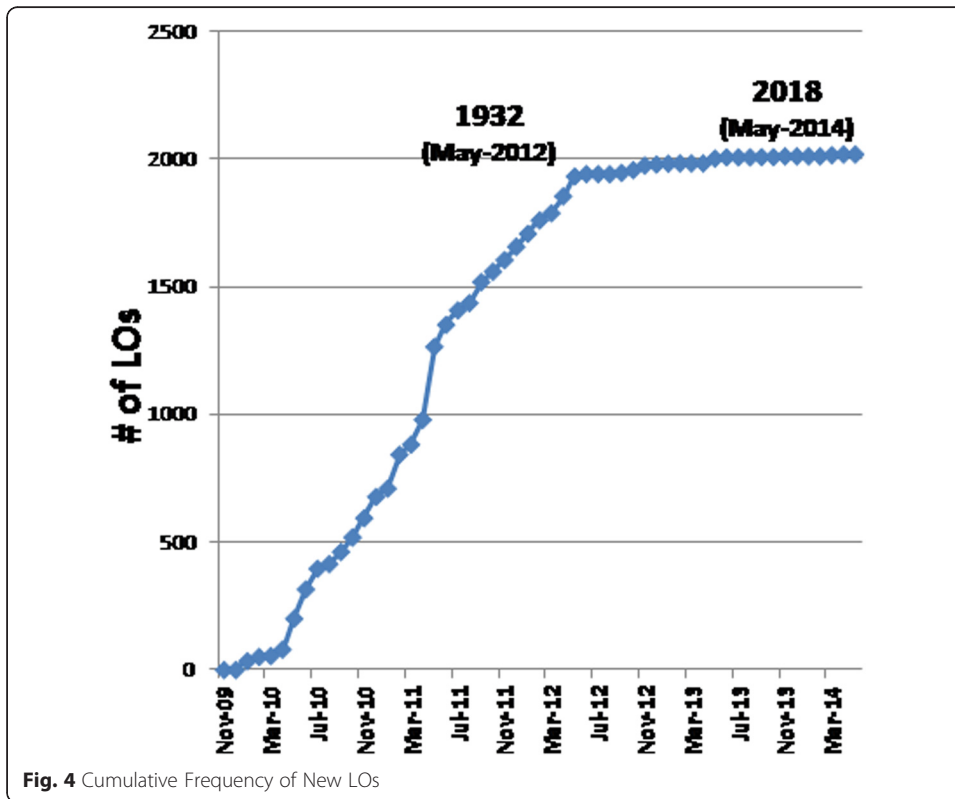
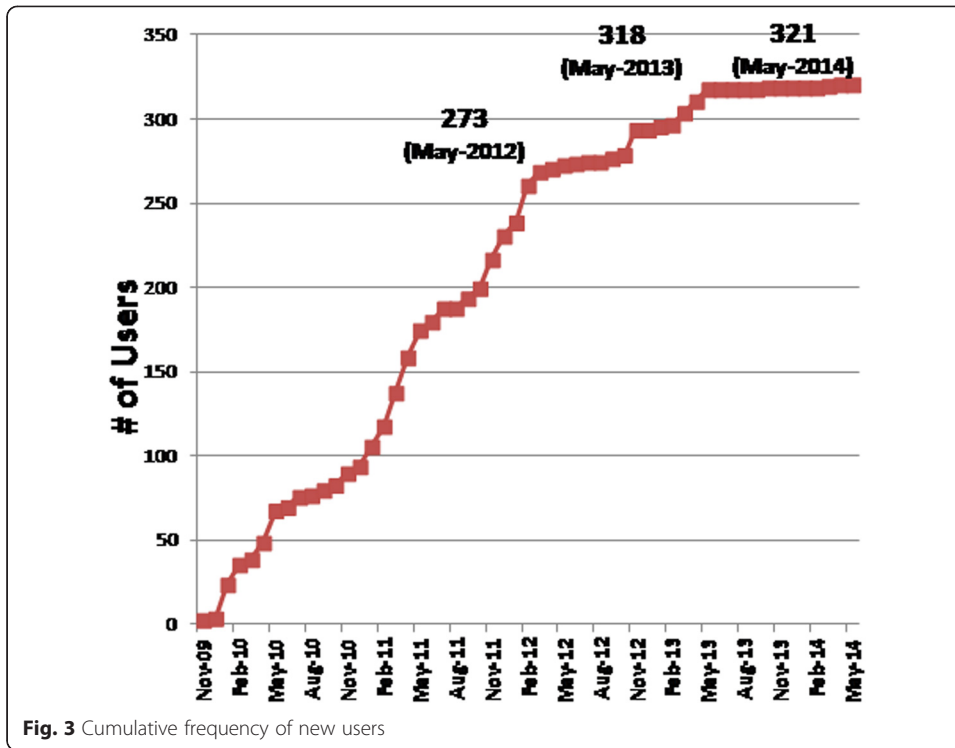


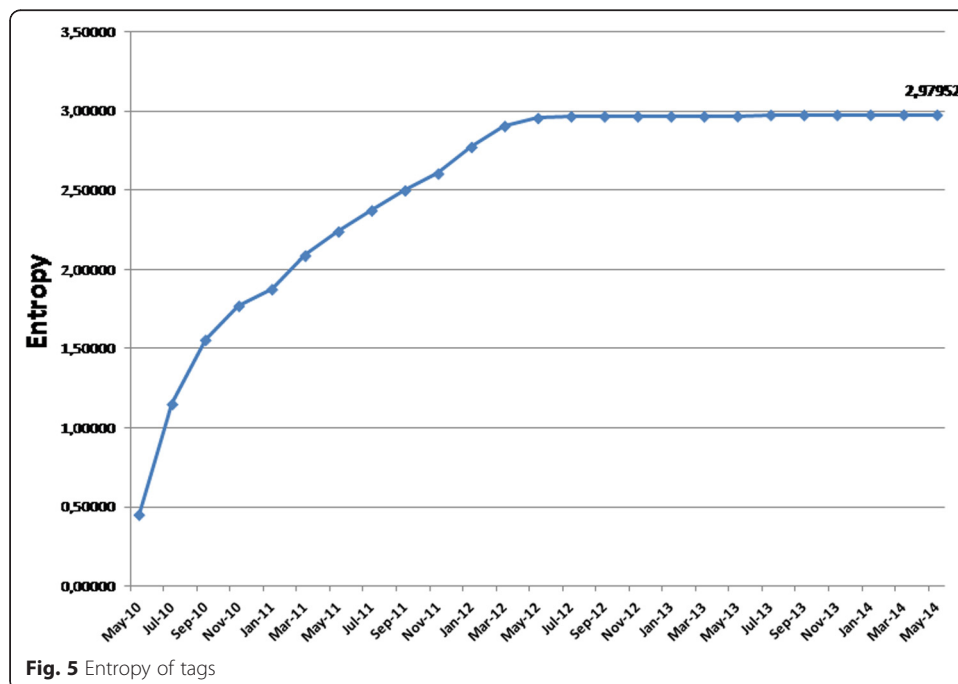
Table 2 Pearson’s correlation coefficient

| | New users | New LOs |
|----------|-------------|-------------|
| New Tags | $r = 0.287$ | $r = 0.545$ |
| | $p < 0.05$ | $p < 0.001$ |

spectrum of distinct tags, which can be used for describing the content of a specific set of LOs does not vary significantly, since science education resources rely on fairly standard and commonly accepted vocabularies across European curricula and at different levels of school education (primary, secondary). Thus, after a certain point, new users can only slightly contribute to the creation of new tags. On the other hand, the addition of new LOs (especially in new subject areas) stimulates the users of the OSR Repository to add new tags for classifying the newly added LOs, contributing to further tag vocabulary growth.

In order to analyze further the trend of the tag vocabulary growth, we have calculated the entropy of tags over time following the formula (1) described in Methodology. Figure 5 presents how the entropy of tags increases over time.

Based on Fig. 5, we can observe that tag entropy follows exactly the same trend line with tag growth. The fact that the entropy line is increasing (until a stabilization point of 2,97952) means that the overall specificity of any tag in the system is being reduced. Furthermore, this also means that tag entropy is strongly related only to the addition of new tags to the OSR repository. Thus, this result provides us with an initial insight that the users are not re-using tags at a high rate, because if this was the case, then it would eventually lead the probability distribution to become less uniform (i.e. entropy will be decreasing) or more uniform (i.e. entropy will be increasing). However, this initial insight need to be validated based on the values of the tag reuse metric that is discussed in Analysis of tag reuse.



Finally, as depicted in Table 3, we calculated the tag growth per LO type following the formula (2) presented in Methodology. The LO types values (presented in Table 3) follow the values proposed by the IEEE LOM standard (IEEE LTSC, 2005).

Based on the results of Table 3, we can notice that LO types with higher interactivity and semantic density such as simulations, videos and lesson plans achieved higher tag growth rates (namely each LO was assigned more tags) compared to other LO types with low interactivity and low semantic density such as texts, questionnaires and images. These results could be useful for designers and/or administrators when populating existing or future LORs, since they provide initial evidence that specific LO types can achieve higher tag growth rates than others. Therefore, incorporating such LO could lead to shorter time frame for the maturing of the tag vocabulary and eventually to its adoption as a supplement to the formal classification system used by the LOR.

Analysis of tag reuse

In order to analyze the tag reuse of the OSR Repository, we calculated the tag reuse following formula (3) described in Methodology and we monitored its value over time. Figure 6 presents how the tag reuse metric changes over time.

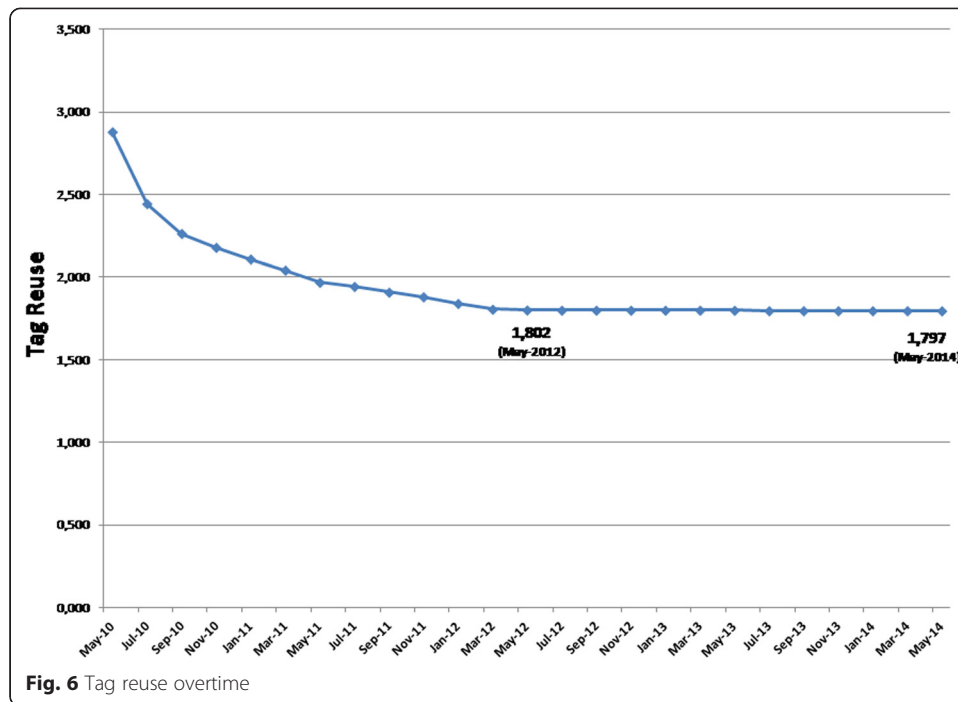
As we can notice from Fig. 6, there is a continuous decrease of the tag reuse metric. This means that the users of the OSR Repository tend to generate new tags to characterize LOs instead of re-using existing tags. This is consistent with our initial insight revealed from the observation of the tag entropy in Analysis of tag growth. The value of the tag reuse metric has been stabilized to 1,797 users/tag. This value is higher than the reported by Farooq et al. (2007) value in CiteULike (1,59 users/tag) and the reported by Vuorikari & Ochoa (2009) value (1,22 users/tags) in Calibrate Portal but still quite low if compared to the reported by Makani & Spiteri (2010) value in CiteULike knowledge management community (23 users/tag).

Moreover, by combining the tag reuse metric with the tag growth metric (as discussed in Analysis of tag growth), we can identify two main periods of the OSR social tagging system based on its behavior, as follow:

- Period 1 (From May-2010 to May-2012): during this period the system had *high tag growth* and *low tag reuse*. This means that the specificity of tags was increasing and this facilitated the navigating to LOs via social tags in the OSR repository.
- Period 2 (From June-2012 to May-2014): during this period the system had *low tag growth* and *low tag reuse*. This means that the tagging behavior was on decline by

Table 3 Tag growth per LO type

| LO Type | # of Tags | # of LOs | Tag growth |
|---------------|-----------|----------|------------|
| Simulation | 615 | 249 | 2.47 |
| Video | 333 | 139 | 2.40 |
| Lesson plan | 216 | 139 | 1.55 |
| Text | 318 | 271 | 1.17 |
| Questionnaire | 179 | 167 | 1.07 |
| Image | 1074 | 1053 | 1.02 |
| Total | 2735 | 2018 | |



the repository users, which could be related to external factors that had to do with the support of the operation of the OSR repository by its owners.

Moreover, it's worth mentioning that the decreasing value of tag reuse could be related to the tagging interface, which does not highly support tag reuse since users are presented (during the tagging process) first with their personal tags and then with the popular tags that has been already added by other users.

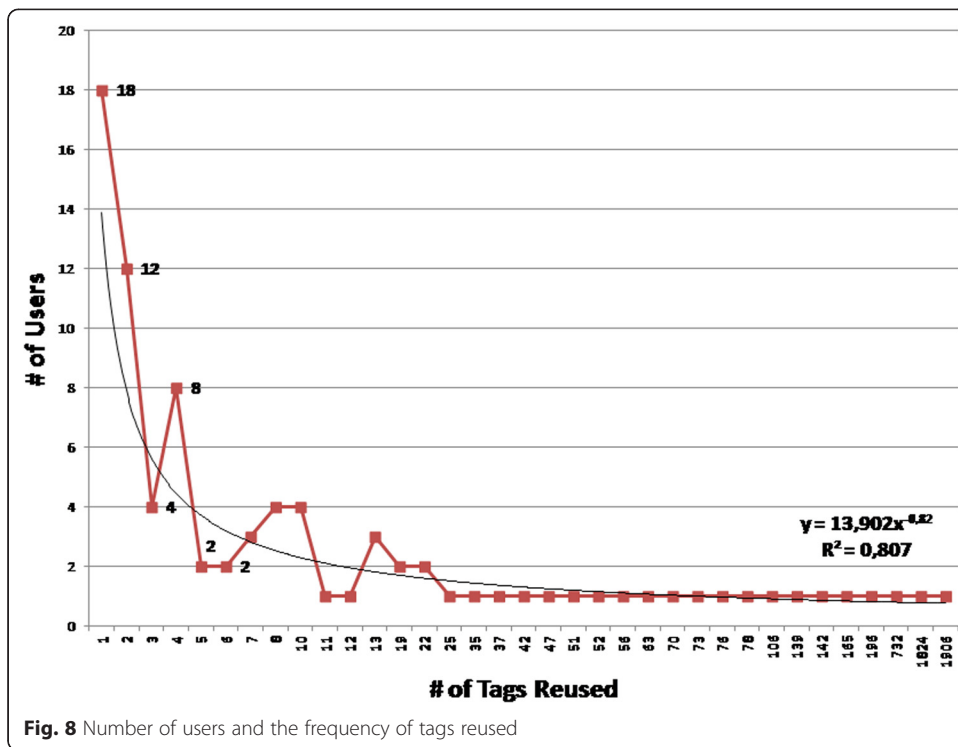
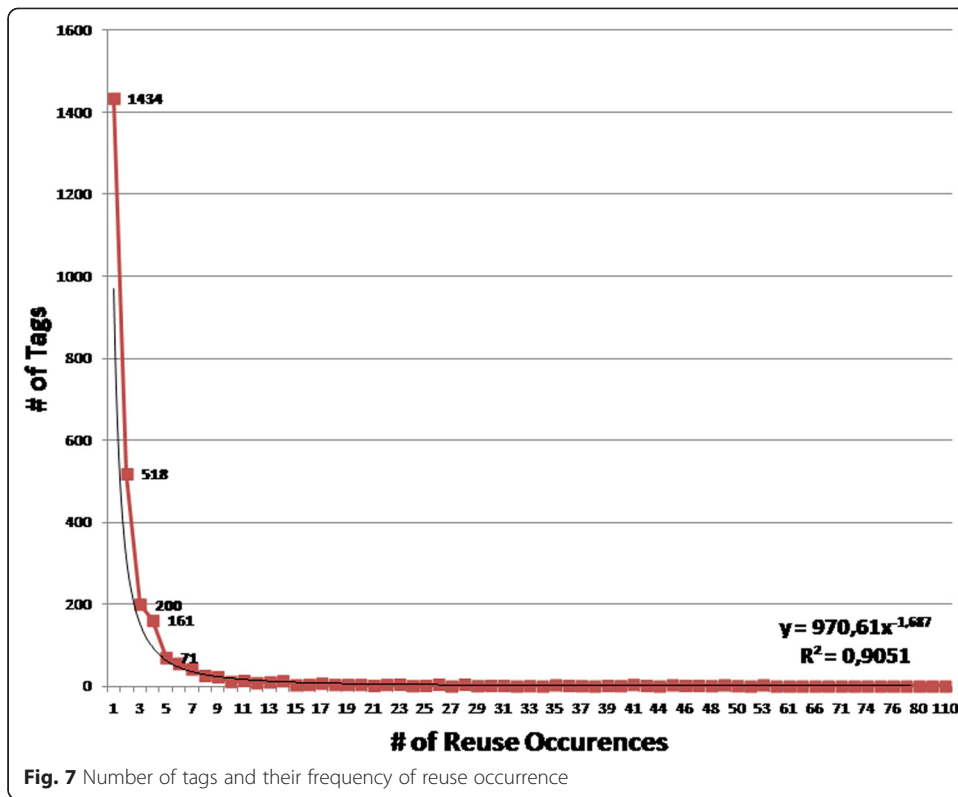
The next step was to calculate the tag reuse per different LO types. Table 4 presents the calculated tag reuse metric per LO type.

Based on the results of Table 4, we can notice that there are not significant differences to the tag reuse metric among the different LO types, since the data revealed a similar tag reuse metric value for all LO types. Thus, we can conclude that for our case the tag reuse metric is not influenced by the different LO types included in the OSR Repository.

Finally, we plotted the distribution of tags' reuse occurrences per number of tags (see Fig. 7), as well as the distribution of tags reuse occurrences per number of users (see Fig. 8). Figure 7 demonstrates a long-tail scheme, namely there are many

Table 4 Tag reuse per LO type

| LO type | tag reuse |
|---------------|-----------|
| Simulation | 1.87 |
| Video | 1.92 |
| Lesson plan | 1.76 |
| Text | 1.69 |
| Questionnaire | 1.75 |
| Image | 1.95 |



tags which have been reused few times but only a small set of tags which have been reused many times. These findings indicate that there are a set of popular tags in the system, which users tend to re-use but since the tag reuse metric is decreasing, the overall distribution of reuse occurrences does not change. This finding is also aligned with the calculated tag entropy discussed in Analysis of tag growth. Figure 8 further corroborates the previous finding, since it demonstrates that the vast majority of users are re-using a very small set of tags. However, there are some “super users” that have re-used many tags (e.g. there are two users who have re-used 1.906 tags and 1.824 tags correspondingly). These findings are also fully aligned with the calculated tag reuse metric. Moreover, it should be mentioned that the distribution of Fig. 7 follows a power law that fits $y = 970,61x^{-1,687}$ with coefficient of determination $R^2 = 0,9051$, whereas the distribution of Fig. 8 follows also a power law distribution that fits $y = 13,902x^{-0,82}$ with coefficient of determination $R^2 = 0,807$. These distributions appear to be similar with distributions from previous studies reported in Robu et al. (2009), Cattuto et al. (2007) and Farooq et al. (2007).

Analysis of tag discrimination

In order to analyze the tag discrimination of the OSR Repository, we calculated the tag discrimination following the formula (4) described in Methodology and we monitored the evolution of its value over time. Figure 9 presents how the tag discrimination metric changes over time.

Figure 9 demonstrates a continuous decrease of the tag discrimination metric, meaning that, overtime, the tags’ capacity to differentiate each LO in the system from the rest, tends to reduce. This finding can be explained since the tag growth metric keeps increasing at a high rate and the tag reuse metric is decreasing. As the Fig. 9 depicts, the value of the tag discrimination metric for the OSR Repository has been stabilized to 3,65 LOs/tag. This value is lower than the reported by Farooq et al. (2007) value in

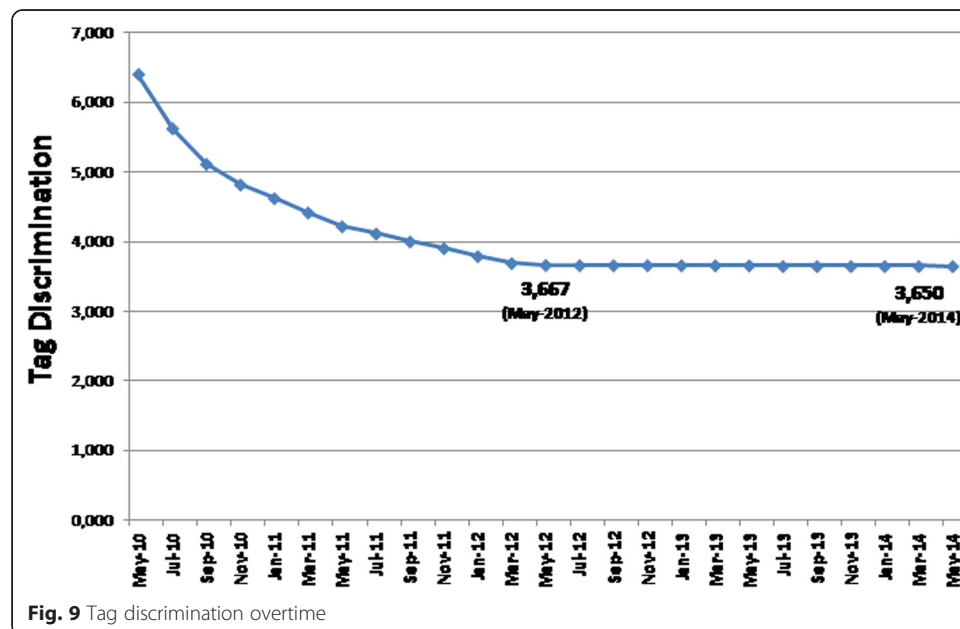


Fig. 9 Tag discrimination overtime

CiteULike (4,47LOs /tags) and the reported by Makani & Spiteri (2010) value (4,11LOs/tags) in Calibrate Portal.

Next, we calculated the tag discrimination per different LO types. Table 5 presents the calculated tag discrimination metric per LO type.

Based on the results of Table 5, we can notice that there are not significant differences to the tag discrimination metric among the different LO types. Thus, we can conclude that for our case the tag discrimination metric is not influenced by the different LO types included in the OSR Repository.

Conclusions and future work

This paper focused on the under-researched area of studying the behavior of social tagging systems within LORs and provided evidence on two major aspects which were not explicitly studied in existing works:

- (a) Analyzing the tag vocabulary of a specific LOR by applying a wide range of tag metrics. The paper used the OSR Repository as a case study and combined the results of the tag metrics in order to generate deeper insights about the tagging behavior of the social tagging system users and
- (b) Perform a more granulated investigation of the evolution of the tag vocabulary, in terms of different LO types accommodated in the LOR.

A summary of the main findings of this study is the following:

- The growth of the tag vocabulary is strongly correlated with the addition of new LOs in the OSR Repository, whereas the correlation with the registration of new users is weak. These findings can be explained considering the focus of the OSR Repository to Science Education LOs. More specifically, the tag vocabulary is expected to grow significantly as new LOs enter the system and teachers can share their insights and experiences on these new resources. On the contrary, the tag vocabulary is expected to grow to a lesser rate when an increasing number of teachers share their (possibly overlapping) insights and experiences on the same pool of LOs.
- Tag reuse in the OSR Repository is mainly focused to support classification of LOs towards future retrieval. On the other hand, reuse of tags for characterizing different LOs towards facilitating the creation of enhanced LOs descriptions is limited. A possible reason for that could be the tagging interface, which does not highly facilitate tag reuse, since users are presented (during the tagging process)

Table 5 Tag discrimination per LO type

| LO Type | Tag discrimination |
|---------------|--------------------|
| Simulation | 3.48 |
| Video | 3.69 |
| Lesson plan | 3.92 |
| Text | 3.31 |
| Questionnaire | 3.53 |
| Image | 3.24 |

first with their personal tags and then with the popular tags that have been already added by other users.

The evolution of tag vocabulary in terms of tag growth was higher for LO types with higher interactivity and semantic density (such as simulations, videos and lesson plans) compared to other LO types with low interactivity and low semantic density (such as texts, questionnaires and images). This means that LOs with higher interactivity and semantic density tended to attract more (distinct) tags from teachers, perhaps due to increased use of such LOs in the everyday practice. On the other hand, no significant differences were identified for the tag reuse and discrimination metrics among the different LO types.

- Overall, the frequency of tag reuse in the OSR Repository is not uniform. More specifically, there are few tags that have been reused many times and many tags that have been reused few times. The same also applies for users, namely there are few users that have re-used many times and many users that have reused few tags. Both distributions of tags per their frequency of reuse occurrence and the users per their frequency of tags reused resemble a power law. This behavior is fully aligned with the behavior of other social tagging systems applied in repositories beyond LORs.

The practical implications of our findings could be useful for administrators and developers of current and future LORs, as follows:

- LORs administrators could monitor the tag growth metric, so as to be able to understand when the tag vocabulary matures and could be used to supplement and/or complete the existing 'official' classification system (such as the IEEE LOM standard) of a LOR. Moreover, by monitoring the entropy of tag vocabulary, as well as the tag reuse and tag discrimination metrics, LORs administrators can understand the tagging behavior of the users of the LOR. These metrics could also be used as a means for providing personalized services to teachers, since they could feed recommendations of LOs that either attract a large number of tags ('popular' Los) or have been tagged by peers with similar past tagging behavior (Klašnja-Milićević et al., 2015).
- LORs developers can develop appropriate tagging interfaces, in order to facilitate the anticipated use of a social tagging system. For example, by providing users with access (during the tagging process) to the popular tags of the system, as well as to the popular tags for a specific LO could facilitate reuse.

Finally, future work could focus on addressing some of the limitations of this paper and provide further evidence on the largely under-researched area of tag vocabulary evolution in social tagging systems in LORs. More specifically, future research could focus on studying the behavior of social tagging systems and tag vocabulary evolution in additional LORs (beyond the OSR repository) with large sets of tags, using the extended set of metrics adopted in this paper. In this way, the insights of this work could be further validated and corroborated with new results from more LORs. Furthermore, future work could also focus on studying the behavior of social tagging systems and tag vocabulary evolution in LORs that are not specific to a particular subject domain (as OSR Repository was Science-specific). This will allow to study the behavior (and the

corresponding tag vocabulary evolution) of social tagging systems in LORs that include practitioners (teachers) from diverse subject domains, and investigate potential differences between them due to this user and LO diversity.

Endnote

¹Calibrate Portal was one of the first European LOR with digital educational resources for School Education.

Availability of data and materials

The dataset(s) supporting the conclusions of this article is(are) available in the OSR repository, [<http://www.osrportal.eu/>] created in the Open Science project [<http://www.openscienceresources.eu/>].

Authors' contributions

Authors PZ and DGS contributed in the design and implementation of the research. PZ, DGS and LP contributed in the analysis of the results and the write up of the manuscript. Both authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Received: 21 May 2016 Accepted: 22 July 2016

Published online: 29 July 2016

References

- H S Al-Khalifa, H C Davis, Replacing the monolithic LOM: A folksonomic approach, *Proceedings of the IEEE International Conference on Advanced Learning Technologies (ICALT 2007)*, 2007, pp. 665 – 669, IEEE
- C Bateman, G Brooks, T McCalla, P Brusilovsky, Applying collaborative tagging to e-learning. *Proceedings of the 16th International World Wide Web Conference (WWW2007)*, 2007
- B Bi, L Shang, B Kao, Collaborative Resource Discovery in Social Tagging Systems. *Proceedings of the 18th ACM Conference on Information and Knowledge Management (2009)*, pp. 1919-1922 ACM.
- Calibrate website. <http://calibrate.eun.org/>. Accessed 17 May 2016
- Y. Cao, D. Kovachev, R. Klamma, M. Jarke, R.W. Lau, Tagging diversity in personal learning environments. *J. Comput. Educ.* **2**(1), 93–121 (2015)
- C Cattuto, A Baldassarri, V Servedio, V Loreto, (2007). Vocabulary growth in collaborative tagging systems, ArXiv e-prints. Retrieved from <http://arxiv.org/abs/0704.3316> Accessed 17 May 2016
- E.H. Chi, T. Mytkowicz, *Understanding the efficiency of social tagging systems using information theory*. Proceedings of the Nineteenth ACM Conference on Hypertext and Hypermedia (ACM Press, New York, 2008), pp. 81–88
- C.W. Cho, T.K. Yeh, S.W. Cheng, C.Y. Chang, A Social Tagging System for Online Learning Objects. *Adv. Sci. Lett.* **4**(11-12), 3362–3365 (2011)
- CiteULike website. <http://www.citeulike.org>. Accessed 17 May 2016
- D. Dahl, G. Vossen, Evolution of learning folksonomies: Social Tagging in e-learning repositories. *Technol. Enhanc. Learn.* **1**(2), 35–46 (2008)
- Delicious website. <https://delicious.com/>. Accessed 17 May 2016
- M. Derntl, T. Hampel, R. Motschnig-Pitrik, T. Pitner, Inclusive social tagging and its support in Web 2.0 services. *Comput. Hum. Behav.* **27**(4), 1460–1466 (2011)
- I. Doush, Annotations, Collaborative Tagging, and Searching Mathematic in E-Learning. *Int. J. Adv. Comput. Sci. Appl.* **2**(4), 30–39 (2011)
- U.D. Ehlers, Extending the Territory: From Open Educational Resources to Open Educational Practices. *J. Open Flex. Distance Learn.* **15**(2), 1–10 (2011)
- U. Farooq, Y. Song, J.M. Carroll, C.L. Giles, Social Bookmarking for Scholarly Digital Libraries. *IEEE Internet Comput.* **11**(6), 29–35 (2007)
- Flickr website. <https://flickr.com/>. Accessed 17 May 2016
- S. Golder, B.A. Huberman, The structure of collaborative tagging systems'. *J. Inf. Sci.* **32**(2), 198–208 (2006)
- T Hammond, T Hannay, B Lund, J Scott, Social bookmarking tools (I) a general review. *D-lib Magazine*, **2**(4), 2005.
- P Heymann, G Koutrika, H Molina, Can Social Bookmarking Improve Web Search?. *Proceedings of the 1st International Conference on Web Search and Data Mining (WSDM 2008)*, 2008, pp. 195-205, Palo Alto, USA.
- Y.C. Hsu, Y.H. Ching, B.L. Grabowski, Web 2.0 applications and practices for learning through collaboration, in *Handbook of research on educational communications and technology*, ed. by J.M. Spector, M.D. Merrill, J. Elen, M.J. Bishop (Springer, New York, 2014), pp. 747–758
- Y.M. Huang, Y.M. Huang, C.H. Liu, C.C. Tsai, Applying social tagging to manage cognitive load in a Web 2.0 self-learning environment. *Interact. Learn. Environ.* **21**(3), 273–289 (2011)
- K. Hyon, A personalized recommendation method using a tagging ontology for a social e-learning system, in *Intelligent Information and Database Systems, volume 6591 of Lecture Notes in Computer Science*, ed. by N. Nguyen, C.-G. Kim, A. Janiak (Springer, Berlin, Heidelberg, 2011), pp. 357–366
- IEEE Learning Technology Standards Committee (LTSC), 2005. *Final Standard for Learning Object Metadata*, IEEE Learning Technology Standards Committee. Retrieved from: <http://ltsc.ieee.org/wg12/>. Accessed 17 May 2016
- A.A. Kardan, M.F. Sani, S. Modaberi, Implicit learner assessment based on semantic relevance of tags. *Comput. Hum. Behav.* **55**, 743–749 (2016)

- A. Klačnja-Milićević, M. Ivanović, A. Nanopoulos, Recommender systems in e-learning environments: a survey of the state-of-the-art and possible extensions. *Artif. Intell. Rev.* **44**(4), 571–604 (2015)
- J. Ma, The sustainability and stabilization of tag vocabulary in CiteULike: An empirical study of collaborative tagging. *Online Inf. Rev.* **36**(5), 655–674 (2012)
- J. Makani, L.F. Spiteri, The dynamics of collaborative tagging: An analysis of tag vocabulary application in knowledge representation, discovery and retrieval. *J. Inf. Knowl. Manag.* **9**(2), 93–103 (2010)
- C. Marlow, M. Naaman, D. Boyd, M. Davis, HT06, tagging paper, taxonomy, Flickr, academic article, to read. in *Proceedings of the seventeenth conference on Hypertext and hypermedia* (2006) pp. 31–40. ACM.
- McGreal, R. (2008). A typology of learning object repositories. In: H.H. Adelsberger, Kinshuk, J. M. Pawlovski and D. Sampson, eds. *International Handbook on Information Technologies for Education and Training*, 5–18. 2nd Edition, Springer.
- OpenScienceResources project website. <http://www.openscienceresources.eu/>. Accessed 17 May 2016
- OpenScienceResources repository. <http://www.osrportal.eu/>. Accessed 17 May 2016
- Robu, V., Halpin, H. and Shepherd, H. (2009), "Emergence of consensus and shared vocabulary in collaborative tagging systems", *ACM Transactions on the Web*, **3**(4), 14:1–14:34
- D. Sampson, P. Zervas, A. Kalamatianos, *ASK-LOST 2.0: A Web-based Tool for Social Tagging Digital Educational Resources in Learning Environments*. In B. White, I. King, & P. Tsang, (Eds.), *Social Media Tools and Platforms in Learning Environments: Present and Future*. (Springer, U.S.A., 2011a)
- D. Sampson, P. Zervas, S. Sotiriou, Science Education Resources Supported with Educational Metadata: The Case of the OpenScienceResources Web Repository. *Adv. Sci. Lett. Special Issue Technol-Enhanc. Sci. Educ.* **4**(11/12), 3353–3361 (2011b)
- E Santos-Neto, D Condon, N Andrade, A Iamnitchi, M Ripeanu, Reuse, temporal dynamics, interest sharing, and collaboration in social tagging systems. *First Monday*, 2013 **19**(7)
- C.E. Shannon, A mathematical theory of communication. *ACM SIGMOBILE Mobile. Comput. Commun. Review* **5**(1), 3–55 (2001)
- G. Smith, *Tagging: People-powered Metadata for the Social Web* (New Riders Publishing, Berkeley, 2008)
- N. Smith, M. Van Coillie, E. Duval, Guidelines and support for building Application profiles in e-learning, in *CEN/ISSS WS/LT Learning Technologies Workshop CWA*, ed. by N. Smith, M. Van Coillie, E. Duval (CEN Workshop Agreements, Brussels, 2006), pp. 1–26
- M. Strohmaier, C. Körner, R. Kern, Understanding why users tag: A survey of tagging motivation literature and results from an empirical study. *Web Semant. Sci. Serv. Agents World Wide Web* **17**, 1–11 (2012)
- J. Trant, Studying social tagging and folksonomy: a review and framework. *J. Digit. Inf.* **10**(1), 1–44 (2009). Retrieved from <https://goo.gl/Zh11ik>
- R Vuorikari, X Ochoa, Exploratory analysis of the main characteristics of tags and tagging of educational resources in a multi-lingual context, *J. Digit. Inf.* **10**(2) (2009)
- R. Vuorikari, H. Poldoja, R. Koper, Comparison of Tagging in an Educational Context - Any Chances of Interplay? *Int. J. Technol. Enhanc. Learn.* **2**(1/2), 111–131 (2010)
- D.A. Wiley, *The instructional use of learning objects* (Association for Educational Communications and Technology, Bloomington, 2002)
- P. Zervas, D.G. Sampson, The effect of users' tagging motivation on the enlargement of digital educational resources metadata. *Comput. Hum. Behav.* **32**, 292–300 (2014)
- P. Zervas, C. Alifragkis, D.G. Sampson, A quantitative analysis of learning object repositories as knowledge management systems. *Knowledge Manag. E-Learn.* **6**(2), 56–170 (2014)

Submit your manuscript to a SpringerOpen[®] journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Immediate publication on acceptance
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com
