# Improving Flickr discovery through Wikipedias

Federico Gobbo

Dipartimento di Informatica e Comunicazione
Università degli Studi dell'Insubria
via Mazzini 5, IT-21100, Varese, Italy
federico.gobbo@uninsubria.it

**Abstract.** This paper explores how to discover unexpected information in existing folksonomies (serendipity) using extensive multilingual open source repositories as the underlying knowledge base, overcoming linguistic barriers at the same time. A web application called Flickrpedia is given as a practical example, using Flickr as the folksonomy and diverse natural language Wikipedias as the knowledge base.

**Key words:** Flickr, Wikipedia, Multilingualism, Folksonomies, Serendipity

## 1 Introduction

Adding meaningful metadata to web content, in order to increase the utility of information by improve the precision of information retrieval to search engines, is one of the most desired feature by any user. Folksonomies are a tentative effort toward this goal. The term 'folksonomy' is a fusion of 'folks' and 'taxonomy' and was originally used in cognitive anthropology studies, but only very recently it became popular with a specialized meaning [9]. A folksonomy is a taxonomy made by tags or labels, usually single-word metadata attached to online items (documents, photos, videos, etc.), in order to add contextual meaning to the items themselves.

Unlike traditional taxonomies, as for example the Linnaean system used in life sciences, there is no explicit hierarchy between tags nor tags are exclusive – e.g. the photo of a cat may be tagged as 'cat' and 'european' and 'animal', but there is nothing that say that all cats are animals: tags can be seen as common facets of the item itself [6]. While in traditional taxonomies there is a central authority that controls its structure, in the case of folksonomies there is no one [5] – undoubtely this is the main reason why folksonomies are becoming more and more popular among web resource users.

Consequently, each tag has two different scopes at the same time: the user's defined one – *personimy*, [5] – and the social shared meaning – *consensus*, as the wide use of tag suggestion interfaces in web applications suggests. Social

meaning emerges when the distribution of tag use converges on some terms, and the distribution curve of tag popularity follows a 'long tail' [4, 1]: very few tags are most used (high consensus, low personimy), and a lot of tags are used once or few times by the majority of users (low consensus, high personimy).

Furthermore, consensus permits *serendipity*, i.e. users dig the web through tags finding new, unexpected and useful content, not easily accessible via traditional search engines. In fact, tags act as filters, i.e. a query on more tags returns the items tagged with any of the given tags – or with all tags, depending on the application [2]. The purpose of this paper is to improve serendipity allowing people to dig folksonomies regardless of the natural language they master.

## 2 Serendipity and multilingualism

Folksonomies share common problems with traditional taxonomies, due to the fact tags are words, i.e. alphabetical strings meaningful in some natural language. In particular, there is no synonym (different word strings, analogue meaning) nor homograph (identical word string, totally different meaning) control. In fact, there is no restriction to what people can write as a tag, i.e. no controlled language: people can externalize their free word association through tags, which respect their own mental models. Consequently, folksonomies lack in standardization, i.e. different strategies in tag encoding are possibles, as for instance dates (28-03-2008, '2008March3', '3r̂d March 2008' and so on) or in the case of compounds ('nice-cat', 'nice_cat', 'nicecat'), not to mention misspellings, so frequent that tag literacy education was advocated [3].

### 2.1 Folksonomies and the digital linguistic divide

One of the existing problems behind folksonomies not fully explored until now is multilingualism. As anedoctical evidence suggests, every tag is written in a human language and users are inclined to write in the languages they are comfortable in. It is certainly desiderable for a user not comfortable in English or other big language (in terms of presence in the web) to search and find tags using a search engine interface in his or her tongue, while the engine searches the corresponding tags in English and in other major human languages.

To do so, the user needs to specify both the tag looked for and the natural language in which it is written in a special web application, which extracts the pairs language-tags in every available language before passing the tags to the folksonomy search engine. Our claim is, when searching in 20 natural languages at same time some interesting photo will be found, that would be undiscovered through a single language search (i.e., serendipity improves).

### 2.2 Adding multilingualism to Flickr through Wikipedias

Flickr, a Yahoo! company, is one of the most popular online photo web applications – e.g., more than 2 million photos are found if 'flowers' are searched, at 2007, April the 11$^{th}$. In Flickr, users can browse or search photos through tags, a feature that certainly contributed to its popularity. Moreover, some open source APIs are available[1] and people can make queries to the Flickr repository through an authentication key given on request. For our application, the language of choice for the API is Ruby, and the development framework is Ruby on Rails, as it is easy to produce clean code and reliable web application very quickly [7, 8].

In our prototypical web application, *Flickrpedia* (named derived from 'Flickr' and 'Wikipedia'), users can make queries in Flickr writing a tag specifying its natural language. The system crawls the Wikipedia in the corresponding language and look for an appropriate page. For example, if the user is a German-speaker and he is fond of airplanes, he may put the following pair `German:Flugzeug` and the system, which can manage case-sensivity, will look for the following page in the German Wikipedia:

```
http://de.wikipedia.org/wiki/Flugzeug
```

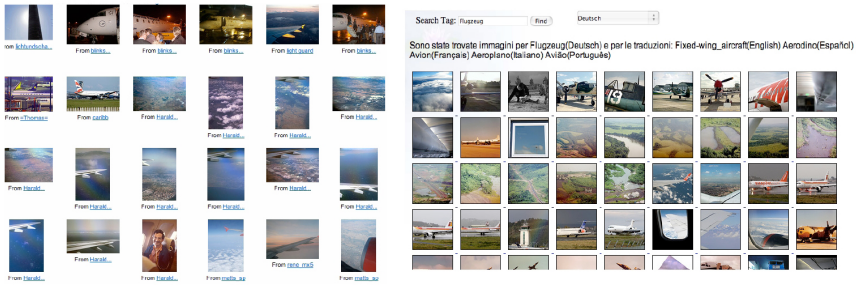


**Fig. 1.** The same query on Flick (on the left) and Flickrpedia (on the right).

where `de` is the natural language ISO code and `Flugzeug` indicates the corresponding web page. With the help of regular expressions, Flickrpedia parses the web page and extracts the existing language pairs of the same topic (airplanes) in other languages from the appropriate web page box known as "in other languages", e.g. `English:Airplane`, `French:Avion` – but also minority languages, as `Basque:Hegazkin` (see Fig. 1). The topic names are passed to Flickr as search queries and thumbnails are given to the user.

While Flickr finds less than 10,000 photos (2007, April the 11$^{th}$) for the tag 'flugzeug' Flickrpedia finds more than 20,000 for the same query, giving a lot of unexpected and relevant photos.

[1] See `http://www.flickr.com/services/api`.

## 3 Conclusions and further directions

This paper has shown that serendipity in Flickr can be improved through the exploitation of Wikipedias's URLs as translation sources. The main advantage is that Flickrpedia should only store the wikipedias according to the existing natural languages – actually, 85. This approach wants to suggest that large and extemporaneus shared information repositories, like Flickr, can be managed through other semi-structured information repositories as the wikipedias – as known, wikipedias are the result of a wide and magmatic community of contributors, even anonymous. Moreover, Flickrpedia, if refined out of its actual prototypical phase, may help users with poor knowledge of major languages to retrieve information only through their lesser-used languages.

Flickrpedia is far from perfect: homographies are still unmanaged, even if wikipedias have disambiguating pages, and it is not clear which wikipedias to choose in order to optimize serendipity. By the moment, the parsed wikipedias are the biggest ones in terms of wiki pages, but this doesn't give any guarantee of serendipity augmentation. Finally, the API given by Flickr is a severe limit: up to 20 tags can be inserted in a single query request, and up to 60 thumbnails may be given.

However, this approach isn't limited to Flickr as the underlying folksonomy. Our research direction is towards generalization, i.e. users can choose the appropriate folksonomy performing multilingual queries. Finally, specific and precise metrics for serendipity are needed, in order to achieve more formally sound results.

## References


1. C. Anderson. *The Long Tail*. The Random House Group, 2006.
2. S.A. Golder and B.A. Huberman. Usage patterns of collaborative tagging systems. *Journal of Information Science*, 32(2), December 2006.
3. M. Guy and E. Tonkin. Folksonomies: Tidying up tags? *D-Lib*, 12(1), January 2006.
4. A. Mathes. Folksonomies: Cooperative classification and communication through shared metadata. December 2004.
5. E. Quintarelli. Folksonomies: power to the people. June 2005. ISKO Italy-UniMIB meeting.
6. Patrick Schmitz. Inducing ontology from flickr tags. In *Collaborative Web Tagging Workshop at WWW2006*, May 2006. Edinburgh, Scotland.
7. D. Thomas. *Programming Ruby*. The Pragmatic Programmers, 2005.
8. D. Thomas and D. Heinemeier Hansson. *Agile Web Development with Rails*. The Pragmatic Programmers, 2005.
9. T. Vander Wal. Explaining and showing broad and narrow folksonomies. February 2005.