

Tagging: Can User-Generated Content Improve Our Services?

Katja Šnuderl
Statistical Office of the Republic of Slovenia
katja.snuderl@gov.si

Abstract

A couple of years ago there was a lot of discussion about how to improve search engines on the statistical websites. We are still struggling to make them find all the data our users are looking for.

On the other hand, in the last few years user-generated content on the internet, with impressive growth of Web 2.0 tools and services, introduced not only user-generated content, but also user-defined classification of items. The so called "folksonomy" replaced authoritative taxonomies and is a result of the tagging. In applications like YouTube (video clip database), Flickr (picture database), SlideShare (presentation database), blogs and others, users attach one or more words (tags) to every object in the database. And when another user types the same word in the search window, all items, tagged with this word, appear on the results list.

It takes just one step to move from entering search keywords ourselves, using all of our knowledge, experiences and intuition in order to tailor the search results to user needs, to allowing our own users to enter tags themselves. This step creates a paradigm shift, exactly the same one as has turned Web 2.0 applications into a big success: Users – not producers – control the way they find and use information. By allowing users to enter tags we can actually allow users to help themselves by helping us.

1. Introduction

In the last decade most national statistical institutes (NSI) designated their websites as their main dissemination channel. The NSI website has moved from publishing electronic forms of selected printed publications in the early beginnings, to becoming an exhaustive source of statistical information, structured by contents and products, layered in time and space. There are press releases, output databases, publications and all kinds of different products available on the websites to be searched for and downloaded. The content of each website is constantly growing. And with the growth of information it becomes more and more difficult to structure the content of the website efficiently enough to facilitate users in search and extraction of relevant data.

For historical reasons each NSI defined its own way of structuring data and information on the website. Mostly a breakdown by themes is used, but also breakdowns by product are not so rare - in some cases access to detailed data, available in output databases, is separated from the main website (but partially or mostly integrated in other web pages). Breakdown by themes isn't a common classification of content which would be used worldwide, but is unique in every country. There have been several initiatives to produce a uniform classification of contents, but so far none succeeded.

Providing easy access to statistical data on the internet resulted also in a significant increase of the number of users. As long as only paper publications and very limited access to detailed data were available, users mostly originated from research institutes and governmental institutions. Making all data freely available on the web significantly changed not only the quantity, but also the diversity of users. Many users don't know where exactly to look for data and they often don't know the exact statistical term. Diversity of users makes providing an effective search engine a very challenging task, requiring constant monitoring of search logs and introducing new search keywords wherever necessary.

Basically there are two main principles of how people find information on statistical websites. One, used by regular users who use the website often, is entering the front page of the statistical institute and then navigating through the website structure. The other - far more common - is to use search engines. In many cases external search engines (like Google, Yahoo,...) are those which bring the user directly into a news release, publication or database table - without visiting the first page of the NSI's website at all.

2. Web 2.0 and the paradigm shift

Web 2.0 is a common expression for a group of very versatile web applications which provide easy interaction between users. They can edit articles, share files and communicate with other users. In the last few years these applications demonstrated a big potential in allowing users to actively participate in publishing content on the web. User-generated content quickly grew into the largest source of information on the web. This quick development wasn't really expected or predicted, it was a matter of some innovative experiments. With improvements in technology, making the servers able to store large capacities of data and developing applications which are in the first place simple to use, one thing led to another and web turned into a platform for cooperation. YouTube grew from a play experiment into the largest video database, Flickr into a valuable resource of images, not only private but also for public institutions and Wikipedia into the largest and most popular general reference work on the Internet.

Most of these web applications share the same basic principles. They allow users to edit articles (different Wiki systems) or share files (videos, photos, presentations, documents). Anonymous access usually allows users only to view uploaded information or files, but users who register and log in can also comment, rate, bookmark and upload their own content.

Other applications within the Web 2.0 group are also blogs and social networking sites. From the viewpoint of this paper they are not so relevant, but all Web 2.0 tools are important for changing users' behavior on internet. The so called "Y-generation" has gotten used to interacting and actively participating on internet - only viewing and using provided information is not what they look for.

In the "Web 2.0 world" there is a significant difference between the numbers of users that only view content that was generated by other users or those who also actively participate. The difference is smaller in social networking sites as MySpace, Facebook, Hi5, LinkedIn

and others, since there every user has to register before being able to access any information about other users. But for sharing sites like wikis, blogs and video/photo/file sharing communities it is clear that the number of users just viewing the content is high, number of users commenting and rating is a bit lower and number of users who publish new content is the lowest. It seems low active participation is also in correlation with quality. Exposing all content to other users makes users follow self-censorship principles and lots of interaction which in general leads to improving the overall quality.

3. Tagging and folksonomy

The above mentioned Web 2.0 applications do not only allow users to share content on the web, comment, discuss and rate it, but they also allow users to organize that content. One way of organizing is bookmarking - putting a link to the item into a user's list of favorite links (usually within the application). Bookmarking is always done for personal reasons, so the user puts the link into his own list of links. Organizing content for all users was traditionally done by placing each item into a predefined category. In most applications there still exist predefined content categories, set by application administrators. So when uploading a video to YouTube the user can define a category it belongs to - Autos & Vehicles, Comedy, Education, Entertainment, Music,... But one can only choose one category, even if the uploaded video could belong to two or more.

To avoid two situations - cases where the uploaded file could belong to two or more categories, and cases where user would want to choose a category not available on the list - tagging was introduced. Tags are keywords, entered as additional meta data to each uploaded file - words that describe the content according to author's opinion and experiences. Each user can add unlimited number of tags to describe the uploaded content. In the application for sharing presentations (SlideShare) the tagging goes even further - users can not only tag their own uploaded presentations, but also other users' presentations. And when another user looks for something (video, photo, presentation), the search engine checks titles, categories and tags at the same time.

If we look at some definitions, a tag is a non-hierarchical keyword or term assigned to a piece of information (such as an internet bookmark, digital image, or computer file). This kind of metadata helps describe an item and allows it to be found again by browsing or searching. Tags are chosen informally and personally by the item's creator or by its viewer, depending on the system. On a website in which many users tag many items, this collection of tags becomes a folksonomy. Tags are a "bottom-up" type of classification, compared to hierarchies, which are "top-down". In a traditional hierarchical system (taxonomy), the designer sets out a limited number of terms to use for classification, and there is one correct way to classify each item. In a tagging system, there are an unlimited number of ways to classify an item, and there is no "wrong" choice. Instead of belonging to one category, an item may have several different tags. (Wikipedia, tag)

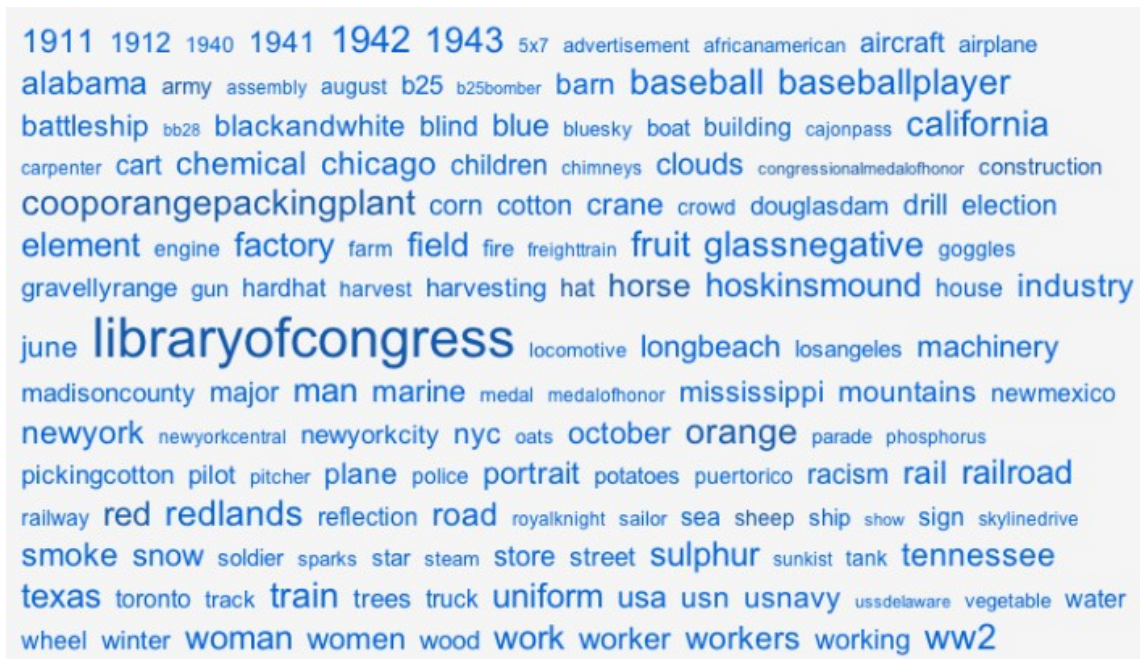
So tagging is a method of categorizing information in a collaborative and decentralized way. Those who participate will categorize information according to their own point of view and agree to share their classification with the other users. The value in this external tagging is derived from people using their own vocabulary and adding explicit meaning,

which may come from inferred understanding of the information/object. People are not so much categorizing as providing a means to connect items (placing hooks) to provide their meaning in their own understanding (Vander Wal, 2007). Once many users enter many tags, the collection of these words creates folksonomy. The term folksonomy is a portmanteau of the words folk (or folks) and taxonomy that specifically refers to subject indexing systems created within Internet communities. Folksonomy has little to do with taxonomy — the latter refers to an ontological, hierarchical way of categorizing, while folksonomy establishes categories (each tag is a category) that are theoretically "equal" to each other (i.e., there is no hierarchy, or parent-child relation between different tags). (Wikipedia, Folksonomy)

Websites where tagging is implemented often display collections of tags as tag clouds. A tag cloud is prepared as a visualization of tag frequency - it's a collection of words, usually sorted alphabetically, where the size of each word represents the number of objects, tagged with this word. Tag clouds also support search, so either you can search for specific words within the cloud or you can search for tagged objects simply by clicking a word in the cloud. A tag cloud can be personalized (collection of tags one user entered), but usually it represents the collection of all tags, entered by all users. A user's tags are useful both to them and to the larger community of the website's uses.

There are some interesting experiences related to tagging and folksonomy. In January 2008 the Library of Congress uploaded over 3000 photos from its collection to Flickr, the photo sharing community. They announced the project and invited users to tag and comment the photos. Decision to launch this project was based on assumption that many users all over the country might know much more about some specific locations and persons on the photos than a few experts at the Library of Congress. When they launched the project, all photos had only one tag applied: LibraryOfCongress.

Picture 1: Tag cloud of the Library Of Congress project after 3 hours



Source: Srivastava, 2008

At the Flickr blog first results were published: "In the 24 hours after we launched, you added over 4,000 unique tags across the collection (about 19,000 tags were added in total, for example, 'Rosie the Riveter' has been added to 10 different photos so far). You left just over 500 comments (most of which were remarkably informative and helpful), and the Library has made a ton of new friends (almost overwhelming the email account at the Library, thanks to all the 'Someone has made you a contact' emails)!" That was after 24 hours. 10 days later the results were: 2440 comments, 570 notes, 13077 unique tags. (Tkachenko, 2008)

There have also been different initiatives to harness collective knowledge and share tags. The aim of these initiatives is to introduce common tagging databases where basic management policy would be introduced and tags could be shared among different websites and web applications. Considering possibilities of semantic web (also called Web 3.0), a few projects aimed to define a common ontology model for tagging. These initiatives are

- SCOT: Social Semantic Cloud Of Tags (<http://scot-project.org/>),
- SIOC: Semantically-Interlinked Online Communities (<http://sioc-project.org/>) and
- SKOS: Simple Knowledge Organization System (<http://www.w3.org/2004/02/skos/>).

4. Introducing tagging in official statistics: SWOT Analysis

As mentioned at the beginning, the websites of statistical offices are rich in content and structure. Users search for data either by navigating through the website structure or by using search engines (external or internal). Now tagging seems to be another option of organizing the content by users, in the way they want and need. Database tables, publications and data release series could be tagged and inter-linked, disrespecting hierarchies and official classifications. For example, a user who works in the corn industry could add a tag "corn" to table on agriculture production, agricultural prices, detailed data for consumer price index (corn bread), external trade, publications on agriculture and anything he finds related to corn. Next time the same or another user types "corn" in the search engine, the search through tags would result in displaying all items that were tagged with "corn" by any user and not regarding the source (output database, file, and news release).

4.1. Strengths

There are several positive aspects with existing tagging systems and the ability to introduce them to statistical websites.

Tagging exists and users know it. The different ways of tagging are already introduced in several web applications and proved to be very efficient. Of course not all users are familiar or comfortable with tagging, but same as for most Web 2.0 applications we can't expect more than a minority of our registered users to participate in entering tags.

Tags lead to intuitive aggregation lists. Once tags are available as simple words with links, most internet users will intuitively recognize these as links and click them to find related

content in an aggregation list. This way they simply start using tags without having to learn about them. Cognitive perception analysis has shown how users in general always notice hyperlinks on web pages. In this aspect tags are easy to use for the majority of users, not only for the advanced ones. But advanced users (e.g. registered users) would be those whom we could invite to participate actively in organizing the content.

Hierarchies are not important when tagging and neither are form/content/status classifications. As an additional metadata layer tags could solve many problems that come from our effort to classify each item into one predefined category. When entering tags, no hierarchy has to be respected. We can enter very general and very detailed expression and both could tell us something about the content of the item. If the item belongs to two different categories, we simply enter two different tags. Also the form is not important, whereas a tag can also describe the form, not just content.

Last but not least, implementing a basic tagging system would require few resources, though it is advised to invest a bit more to build a system which would support also long-term management policy and advanced features.

4.2. Weaknesses

There are also some bad sides to tagging we should be aware of and take care to manage such a system properly in order to avoid them.

Unimaginable to statisticians: In the "statistical world" a lot of communication respects strict coding systems and standard classifications. Tagging is exactly the opposite and can disrespect all expertise. It doesn't mean users who are very advanced wouldn't use their own expertise, but there is no insurance tags will be in any way ordered or "correct", at least at the very beginning. But that is the main characteristics of tagging - different users have different perspectives and there is no such thing as a correct and incorrect tag. It can very likely happen that users enter tags which would never appear on the list of search keywords, entered by statisticians or webmasters.

Difficult to get started: Tagging in general works as the "snow-ball effect", meaning at the beginning it might take time to get users involved, but more tags are added, more users will see their chance to participate. These starting period problems might be managed by entering tags by statisticians and editors at the statistical office, or by inviting registered users on a more personal level to actively participate, in a similar way as it was done in the Flickr project by the Library of Congress.

Unbalanced tags: As the start tends to be difficult, we should also expect that not all content will be tagged at the same grade. If one user tags an item with a certain word, he might not find all related content and tag it the same way. To create a balanced collection of tags for each item on the website, quite some time would be needed. This problem could be partially solved by managing tags by website owners in a way where they would monitor entered tags, analyze users' behavior and add similar tags to all related items. Tags can always be entered by either authors or users, so there is no collision with basic principles of tagging.

Grammatical rules: It can easily happen that a user mistypes a word or that different users use different grammatical forms (singular/plural...). When tagging photos, there is a difference whether the tag says "woman" or "women", but with statistical data this should be related content. A way to avoid this problem is to both monitoring tags and correcting mistakes, and on the other side to use tools which suggest words (from the existing list of tags) while typing.

Whereas introducing a basic tagging option doesn't take much effort, it is more resource-consuming for advanced features. To ensure quality of tags, balanced situation and overall consistence, some managing policy should be introduced. But some website editors already monitor entries in current search engines to improve search keywords on the website; basically these principles are the same.

Multi-linguality: Most statistical institutes provide their websites in both national and English language. But even if the content of these websites is the same, it is not in the nature of tagging to implement automatic translations. Several words can have more than one meaning in another language or another possible interpretation, which might be out of the context. And also the perspective of national and international users might differ. The only way to deal with multilingual pages is to treat tags in different languages separately.

4.3. Opportunities

Thinking about all the opportunities tagging could bring, if we promote it enough to engage our users, we can give our imagination a pair of wings to be carried above the clouds. But some opportunities are also very realistic.

User-centric search support: Tags are by their nature user-centric, since tagging is done by users. There is no need to analyze common expressions used in daily life to introduce them as search keywords, users will enter these words themselves. The search engine can look at both "official" search keywords and "unofficial" tags, bringing the content of our website closer to the user.

Integrate search clouds and tag clouds: Tag cloud is a collection of alphabetically sorted words where which word is presented in a relative size. The difference between search clouds and tag clouds is that in a search cloud the size tells how many times the word is searched for, but with a tag cloud it means the number of items tagged with this word. A common principle to define the word size should be used. Word clouds work as an intuitive search engine as well - users can at once see "what is available" on the website, at a highly aggregated level. Large words draw attention for their popularity, smaller ones for their speciality. Clicking on a word brings the user immediately to the list of related items, where either tags or search keywords were entered.

Get to know our users better: As tagging is done by our users, not only website editors but also statisticians can easily see what kinds of words users attach to related items. Of course there can be some words which seem irrelevant, but there could also be some valuable information and added value in sharing experiences in terminology.

Higher engagement of advanced users: Registered users, who usually represent advanced users, could be personally invited to tag items on our website. Many of them might be willing to actively participate in organizing the content of the website, which would mean real interaction with users. Not just asking them about their opinions on how to organize content better, but to simply allow them to organize it in a way to tailor their own needs. Most active users could also be recognized with a thankful letter or a small gift at the end of the year. Respecting users this way could improve the overall impression of user-friendliness, and advanced users would at the same time help all other users find relevant information easier.

Product (not instance) level tags: Tags can be made re-usable, if they are not applied only to one item of the kind (e.g. press release in one month), but on the product level (e.g. press releases about the same topic can share same tags). This kind of tagging is not very common, but the policy should be clearly presented to users who enter tags. But on the other hand this means that tags should not be describing trends, reflecting one single data release.

Share (English) tags on international level: It is interesting to imagine a tag cloud where English tags from all statistical websites around the world would be collected. With one click the user could find related content for many different countries. But in order to support this option a common platform should be used and a big effort invested. Though the proposed ontology models, which aim to develop a common standard for tag sharing, already suggest a common platform. These should be investigated further in order to support tag sharing among statistical institutions worldwide. If the project succeeds, it could eventually develop into a tag-driven worldwide statistical resource database.

4.4. Threats

There are also a few threats we should be aware of if we introduce tagging. To start with, there could be no interest from users. To avoid this problem we should engage them actively. Where one user leads, others will follow. Therefore it might be wise not to start with empty tag list, but to enter a few tags for each item ourselves and then continue with most advanced and active users. Of course tagging made by website editors or table authors should also be a permanent action in line with tagging policy, to maintain balanced tags.

There might also appear spam tags. When monitoring the tag database, all spam tags should be removed immediately and if necessary also implement anti-spam policy. There are several ways, but none absolutely reliable. It might be wise to enable tagging only for registered users, at least for start.

Some people might also say that allowing our users to describe the content of items on a statistical website with their own words by tagging might jeopardize the "Official label", which is very important for official statistics. But tagging isn't about data, it's about making search for data easier.

6. Literature

1. Ten Bosch, Olav, De Jonge, Edwin: From statistical data supplier to statistical information service provider. International Marketing and Output Database Conference. Avila, 2006.
2. Jenssen, Jesper E.: Searching www.dst.dk or the challenge of the Google myth. International Marketing and Output Database Conference. Den Haag, 2005.
3. Kim H.L. et.al.: Review and Alignment of Tag Ontologies for Semantically-Linked Data in Collaborative Tagging Spaces. URL: http://scot-project.org/pubs/kim_ReviewAlignmentTag.pdf, 15.9.2008
4. Roy, David: Strategic Vision of Internet Presence for National Statistical Offices. International Marketing and Output Database Conference. Den Haag, 2005.
5. SCOT: Social Semantic Cloud Of Tags <http://scot-project.org/>
6. SIOC: Semantically-Interlinked Online Communities <http://sioc-project.org/>
7. SKOS: Simple Knowledge Organization System <http://www.w3.org/2004/02/skos/>
8. Srivastava, Kakul: Tagging - web 2 expo 2008. URL <http://www.slideshare.net/kakul/tagging-web-2-expo-2008>, 24.11.2008
9. Tkachenko's Blog: Crowdsourcing in Action: The Library of Congress Photos on Flickr. URL <http://www.tkachenko.com/blog/archives/000727.html>, 24.11.2008
10. Vander Wal, Thomas: Tagging That Works - O'Reilly Web 2.0 Expo 2007. URL <http://www.slideshare.net/vanderwal/tagging-that-works-oreilly-web-20-expo>, 24.11.2008
11. Folksonomy. Wikipedia, the free encyclopedia. URL <http://en.wikipedia.org/wiki/Folksonomy>, 24.11.2008
12. Tag (metadata). Wikipedia, the free encyclopedia. URL [http://en.wikipedia.org/wiki/Tag_\(metadata\)](http://en.wikipedia.org/wiki/Tag_(metadata)), 24.11.2008
13. Web 2.0. Wikipedia, the free encyclopedia. URL http://en.wikipedia.org/wiki/Web_2.0.