

HUBERTUS KOHLE (LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN)

Social tagging in the humanities and the nature of the internet workflow

Abstract:

In order to collect data on specific objects - as for example artworks - we usually rely on expert contributions. The paper is intended to explain alternative procedures called social tagging. It is worth discussing different aspects of such procedures, which allow for superfast database development, farreaching possibilities of data analysis, and most generally show revolutionary aspects of the workflow in the internet.

This is the text of a lecture given at the conference »Nuovi sguardi – nuove prospettive« at Ca'Foscari University, october 13th, 2011. The form of the spoken text has been retained.

Introduction: The concept

With my paper I want to draw your attention to a field of literary practice which has not often been seen in conjunction with the sciences - let alone with the humanities. The reason for this is to be found in the fact that social tagging is intimately connected with the internet, still a somewhat embattled medium especially in the humanities, and even more importantly in the way data are collected. The two words in the concept of social tagging refer to the activity itself and exactly this way of collecting data just mentioned. To »tag« means to annotate a resource, be that a text, a sound, a picture or whatever; »social« is related to the source of these annotations, which is not the individual expert but a more or less unknown member of the »mass«. To be sure: this tagging process is completely free, the taggers can write down whatever they want, they are not bound by ontologies which coerce them in a controlled conceptual system.

What can the crowd contribute to sublime sciences and humanities? It is known in social bookmarking systems, in facebook, flickr, and a whole range of evaluation sites in the internet. The famous »I like« in facebook will be of help when the problem arises if the restaurant on the other side of the street is worth to be visited or not, but for physics or art history? Aren't they too complicated and mysterious to become controlled by the undefined crowd? Yes and no is my answer, and as you might expect, most of the time I have for this paper will be dedicated to the »no«.

Social tagging in the library

The one sphere where social tagging has had a certain success in academic work and has been taken seriously is the library. Libraries have certain problems with traditional ways of making accessible their sources, which means books. The annotations on the content of the book selected by experts are not identical with those that are taken by the readers when they look for publications. And they tend to be less and less so. This does not mean that the keywords chosen by the experts are bad ones, probably even the contrary is true, which means that those figured by the users are somewhat amateurish. But keywords are made for these users, you better change these words than the readers.

Why not give readers themselves the possibility to add such keywords? I know of some few libraries that are experimenting with such procedures, and you will be able to imagine the huge resistances within these libraries by librarians afraid of losing their field of expertise or even their employment. But apart from the fact most of even the innovators underline that expert and social tagging are not alternatives but mutually supporting procedures, there is one central problem connected with tagging books - or whatever media published in written form. And this is the time lag between looking at the tagging site, which is in this case the library catalogue, and coming back to it when readers are able to annotate the book. You evidently have to read the book before you are able to add keywords to the catalogue, and who will be willing to come back to the site after having read the book?

Social tagging for the visual arts

Another object offers more chances for a successful tagging, and here we are approaching our own field of interest, the one which is related to the history of the arts. Pictures have always been considered inferior to texts in a logocentric world. They are universally understandable but at the same time they are flat - at least this is their reputation. Text needs time in order to be understood, pictures open up immediately - again considering what is generally thought. What is a drawback in this perspective might become an advantage for us who are interested in finding out about the potentials for social tagging. A picture is immediately ready for being annotated, and if it has a sky, there will be many of the taggers who write down »sky«. And they will be right to do so.

This is the basic idea that we are realizing in a project which had very simple aims in the beginning and afterwards demonstrated to have farreaching consequences.

<http://www.artigo.org>

You are all familiar with digital image databases, and you know that some of them are really huge. I just mention »prometheus«, which is a meta-database integrating almost 60 individual databases with about 800.000 reproductions of artworks produced by private enthusiasts, museums, and universitarian institutions. (www.prometheus-bildarchiv.de) As it is a lot less time consuming and even partly automatable to scan the images than to make them textually accessible - and for the time being textual metadata reflect the common academic approach to works of art - every single dataset often includes only very basic descriptions of the work, and they especially lack iconographic and stylistic metadata. This means that the potential of such databases is rarely fully exploited. The traditional way would be to engage experts for annotating the images, using for example »iconclass« for a professional iconographic valorization. It is obvious that this is a very laborious and expensive procedure, and considering the fastly growing image databases it seems to be doomed to fail.

So let's invite the crowds to do that. Okay, they are not checked about their qualifications, but on the other hand, they are many, and they might work for free if you allow for alternative satisfactions. The problem will be to fight the negative and to enhance the positive conditions. Of course we first had to fight the quality reservation which is sort of killer argument against many of the new internet applications. There is one very basic critique which is grounded in the facts - and this is indeed a fact! - that completely unknown internet players might tend to be trolls and intentionally make bullshit annotations. Indeed the use of a tag »Arnold Schwarzenegger« under an image of Saint Peter is dubious. In order to avoid these obstructions we used the ingenious idea of American information scientist Luis van Ahn who organized the annotation process in the form of a game. (<http://www.gwap.com/gwap/>) Two persons connected to the internet were invited to play against each other, and only in case that both tagged the same word, this word was accepted, »matched« and included in the database for future searches. The wisdom behind this idea: We would even then not be able to avoid someone who tagged Saint Peter with »Arnold Schwarzenegger«, and another one with, say, »Barack Obama«, but we could be almost 100% sure that never ever two persons not known to each other would both tag »Schwarzenegger« or respectively »Obama« at the same time.

This is indeed the basic idea, the rest is almost routine, although technically quite demanding. The tagging game was called »Artigo«, we had a proof of concept running online for more than two years. Then we got a grant from the German science foundation in order to produce a more professional version and to increase the number of players. By now we have almost 7000 registered players and 4000 with no account. We have collected as many as 750.000 matches after 3 years, and what might also be interesting: the number of matches

amounts to only 23 % of the tags, which means that there is a lot of nonsense or many, many sophisticated words not yet matched and subject to a future matching process.

As you can see in the top of the introductory screen, we have an English, a French, and a German version of the game, but understandably most of the players are German speaking. (www.artigo.org) Below you can register in order to collect points which you get for successful, which means matched tags. We used to pay 50 Euros for the monthly highscorer, but this has become too expensive, and actually people do not play for the money - but for a mixture of self-interest and cooperativeness. Self-interest because they hope that what they carry out will also be done by others so that soon a powerful search tool will be available. And cooperativeness, because obviously – this seems to be a big topic in economics – the human being is less self-referred than normally conceptualized in our liberal-capitalist world-view. On the right you see the search box. Below in the left middle you get some informations on *artigo*, although we generally try to avoid lengthy explanations in order to rely on the fact that the game explains itself. Of course there is a list of the best scorers because the competitive aspect is very important. And then you see the blog which is an important element in the game because it strengthens its social aspect. In the blog collaborators in the projects give their ideas about general aspects of the game, but also on the whole problem of social tagging, and they of course answer to comments and questions by the users.

As you can also see there is yet another game, and we called it *Karido*. There are two explanations for adding further games – and I hope we will still add at least one further version. The one is »tactical«, the other tries to tackle the basic problem of tag quality. As we do not refer to a scientific, but to an unspecific internet community, the game has to be fun. (By the way: The somewhat populist comments which we added below the pictures on the introductory screen are also intended to increase the »fan-factor« and at the same time to pin the attention of people surfing by and usually going away very quickly) A game in itself should be fun, but games tend to become annoying after a while, and in order not to lose our gamers, we try to switch them to another game. *Karido* is at the same time designed to enhance the quality of the tags, or better, to enhance their specificity: The player gets a whole range of images which tend to be visually similar, and he has to describe one of them in a way that this description can not be confused with another one. Therefore he cannot be generic (blue, woman, horse), but has to be more specific (darkblue, queen, grey horse). When the player is not precise enough, he gets added request from his partner (the »guesser«) to specify his annotation. What I would love to do next: to set up an editorial game which I would call »tag-a-tag«. After three years by now we have a lot of images tagged many, many times, and some of them have 30 to 40 matches. Now what we register is that after having been presented dozens or even hundreds of times to the players, now

and then errors creep in. Simple orthographic errors but also interesting ones, for example when a painting by Giulio Romano is matched with »Raphael«. This is of course interesting, because not only at first sight a Giulio Romano can be very similar to a Raphael. In order to qualify such errors, I would propose to invite people to tag a tag. They will get the image and one of the matches and then they have to decide: Is it a good tag or is it a bad tag. In order to be somewhat more subtle, one could also offer them a scale from 1 to 10. A painting with just some blue spots would only get a »2« for the tag »blue«, an Yves Klein would receive a massive »10«. But the basic function of this game would be to correct obvious errors. Do you recognize the consequences of such procedures? Can you imagine other applications where such approaches can be functional? I do. The tedious transcription of handwritten texts could be delegated to the crowd, just to name an example, and you only have to find ways of giving this a certain fascination in order to seduce people to do something as tedious as this. Because then it is fun. By the way: Do you agree with me that this is not only a way to yield work to the crowds but also to introduce society at large to what we are doing in our respective fields? And that this can be very helpful to justify our work?

If I had more time I would now show you a game in action. You would see, that we have set a time limit in order to enhance the competitive character of the game. You would see the number of tags given by your partner, but of course not the word itself, because then you would just copy it. What is extremely interesting for educational uses is the end of the game. After five minutes of strained tagging all the tagged pictures are presented again and now including their metadata: artist, title, dating. The idea behind is that people tend to remember better after having worked conceptually with a picture – as they have done while tagging it - than just having looked at it. Our art history students all have to have a certain knowledge of canonized artworks. This is admittedly not a very modern approach, but we still think it is important. Most of them scroll books in order to learn these metadata, but I am sure it is a lot more intense when they do it in the context of such a game. Give them 5000 reproductions of the essential of art history in such a game (whatever essential is defined in an age in which the canon does not have good publicity), and they will have memorized them. Once upon a time I was engaged in developing elearning units, and as many others we made the fault to follow quite strictly the teaching models of the real world. Having made the experience with social tagging I would tend to underline that such ways are didactically probably more effective, with the restriction, of course, that a game like artigo is not intended for somewhat more complex discursive reasoning but just for single or few word tagging.

In the way described we already have achieved two aims: We quickly have established an annotated picture database, and we also realized a didactic aim by making people acquainted with artworks. But there are other fascinating perspectives which we will still have

to realize and which I can for the time being only theoretically describe. Before I will do this, I will again discuss a fundamental skepticism as to the quality of the tags. It is true that we avoided the grossest obstructions by organizing the tagging process as a game with two persons the one unknown to the other in the internet. But this does not avoid that we mostly have very simple annotations which do not really match with professional expert descriptions. This is and it remains a problem. But there are ways to improve the situation. Many of them, to be sure. I will just suggest two.

You will remember that more than three quarters of the tags are not matched and remain hidden in the storage of the computer. A lot of them are misspellings, troll tags and so on. But many of them are also extremely high quality, so very high indeed that it takes time to find a second person who matches them. This second person is not found in actual life games but in what we call simulated games. Very often we do not have two real players at the same time, which is due to the fact that 10.000 theoretical players are not sufficient to have two of them at the exactly same time online. But what looks like a problem could also be an advantage. Because every time there are no two players present we confront the one present with an old game which we simulate. The player really doesn't even notice it, because it looks like a real game, and I am not sure if the one I showed you before was not a simulated one. But if two years ago someone tagged with »apocatastasis«, that is the idea of a universal forgiving of the sins at the end of the days mostly present in modern protestant theological thinking, this tag certainly has to wait a while until it is matched. But one day it will be, and the tag waiting in the computer storage is reconciled itself by being matched.

If this sounds too esoteric to you, I can offer another solution, and this solution is assured by a sophisticated gratification procedure. Normally you get a point for a match, but why not give 20 points for an expert tag? As soon as our players notice that they receive 20 points for an advanced concept like for example »apocatastasis«, they will strive to find such advanced concepts. How can we automatically guarantee such a process?

There are several computerlinguistic application which defines the frequency of a word. »blue« and »woman« are very frequent words. »apocatastasis« is very rare. We connect our game to this application which is available online, and we tell the system to allocate a high score for rare words and a low one for frequent words. The assumption is, that a frequent word is a generic one, and that a rare word is a sophisticated one. This will not always be true, but generally it is. So the problem is solved and we have found a way to improve the quality of the words, because we are sure that having played a while people will understand the mechanism and stray for entering more uncommon words.

Tagging as research

All I have been speaking about until now is on collecting many and good tags in order to improve the searchability of a picture database. The real research aspect comes up when we begin to analyze the tags in themselves, and the way they were produced. The idea: Tags report about two different entities: the individuality of the tagger, and the characteristics of the items tagged. In both fields it might be interesting to ask for the concepts chosen and also for the sequence of tags. This last point is feasible, because we give a time stamp to every chosen tag. It is trivial to suppose that different taggers assign different concepts. A child will find other tags than an adult. But there are more exciting and less obvious cases. What about women and men? Or what about Italians and Chinese taggers? It is known and adopted for example in eye-tracking experiments that persons from East Asia look at pictures in another way than Europeans. A European tends to begin with looking at the foreground, someone from China begins with the background. I wouldn't be surprised if such cultural differences are also expressed in the sequence of tags allocated. Certainly it is a fascinating field of study, fascinating also because it opens up traditional art history to more general visual studies which in Germany figure under the fashionable title of »Bildwissenschaften«. And for a more classical study of artistic styles, there could also be rewarding aims of research. Will an impressionist landscape be tagged differently from a classicist one? Not just in the concepts chosen but also in the sequence of their allocation, which would not be surprising because a landscape by Claude Lorrain is organized in a way which guides the eye through the painting differently from one by Monet. Such implicit knowledge could also be found in an example as the following: What will be the nature of a painting tagged with »red«, »diagonal« and »dynamic« as the first three? Will it be a Rubens with a lot of red in it, a typically baroque dynamism produced by diagonals in the painting? I do not think so because taggers will probably first say »man«, »Jesus« and »Mary Magdalene« and maybe later also »red«, »diagonal« and »dynamic«. No, it seems to be a lot more probable to me that it will be an abstract painting. Will a painting tagged a hundred times with blue be »bluer« than one tagged only 5 times? An important problem for perceptual psychologists. How many similar and rewarding problems can be figured as soon as we have understood the nature of such a game!

Tagging in other fields

As I do not have enough time I just mention that such games can of course also be considered in other fields, not just in art history. Imagine the uses in linguistic field studies! To

me such games have the character of a real paradigmatic shift in a lot of empiric branches of the humanities

Tagging and the nature of the internet workflow

If we consider social tagging as a procedure of collecting data in a more general way, something quite revolutionary about typical internet workflows might appear. Traditional knowledge building especially in the humanities relies on solitary work of a single thinker, the one who is working in the famous ivory tower. In the eyes of such a thinker everything has to be deeply reflected about before being written down, the resulting text is the outcome of a spiny act, the dear child of a long time full of privations. To put this in the sober words of information science: The filtering process takes place early, in the brain of the thinker, before anything is put on the paper. By the way: I beg you to understand that there is not even the slightest irony in such declarations, although it might seem to be. For the sake of clarification I just exaggerate the difference to typical internet workflows: Here there is a tendency to prompt huge masses of information, which filtered later in order to enhance their quality. Those masses are taken from the most heterogeneous sources, the whole accent of the procedure is laid on a sort of »post-production«, which assures that the gold is filtered from the mud. »Publish first – filter later«, that is the dictum by Clay Shirky that summarizes such a procedure and which is extremely important as the definition of a typical internet workflow. (Here Comes Everybody: The Power of Organizing Without Organizations, Penguin Press 2008)

It can be transferred to other fields outside social tagging and I would like to finish my reflections by adding some thoughts about them.

Most of you will have made the experience that it is a lot less arduous to begin with writing down a text in the digital medium than it was before when you sat in front of your sheet of paper which looked at you in all its splendid and reproachful emptiness. The ink on the paper had something definite, the letters on the computer screen can be obliterated which gives us the good feeling that even the silliest ideas can be completely forgotten. And if not completely forgotten, then at least they can be worked and reworked and totally transformed, until the silliness has become a wisdom. Something similar can be observed in Wikipedia. Mostly the texts in Wikipedia are the result of a common effort, a current procedure here is that someone produces a short basic text, and as soon as this has been done, others go on with it, as if that had waited for someone making the first step. In my eyes, the relocation of the filtering process ahead in time, postponing it from the beginning in the sphere of pure thought to a later moment is one of the most central and far-reaching aspects of intellectual

production in the internet. If we generalize this insight, it can for example be used as a point of departure for defining the nature of the publication process in the internet age.

Scientific publication

As a fluid medium, the electronic text lacks the hieratic qualities the printed text used to have. This text in its medial form is a work in progress, never definite, always object of constant adaptation. In the printed form, the text is ponderous and not universally addressable, in the digital it can be distributed extremely quickly and multifariously reconfigured. On the negative side, this is often criticised because it seduces to a very speedy text production. But there is also a positive side. More persons than ever before are able to write down their thoughts, and if this is also taken as a problem by highbrow intellectuals, it is because they feel scared in their leadership. Jürgen Habermas and his thoughts about the internet might be a case in point, but you can also take traditional journalism under the pressure of bloggers. (Jürgen Habermas, Preisrede anlässlich der Verleihung des Bruno-Kreisky-Preises für das politische Buch 2005, <http://www.renner-institut.at/download/texte/habermas2006-03-09.pdf>) In scholarship in the stricter sense of the word something similar can be observed, or, to be more precise, foreseen.

In print-culture, we are constrained to implement a prior filtering process in order to separate the wheat from the chaff – and if this was only for the lack of paper. This is usually organized in the form of a peer-review. In the digital world, this is no more necessary, maybe even not recommendable. Because we might find alternative procedures there. One of them could be an evaluating comment post festum, which is of some help for the reader drowning in the oceans of publications. If we add the precise addressability and the possibility to exactly assert the frequency with which a specific document has been retrieved, the problem of selection is half way solved. And maybe better than today, where this problem has become urgent even under analogous conditions. Further electronically based methods for determining the pertinence of a text with regard to one's specific way of posing a problem can be added. For example, I could imagine an automatic matching process with the qualification of an author in the internet.

Even the tiny nibbles of information, which in the eyes of the pessimists tend to dominate in modern publication culture in order to blow up one's record of publications, then appear in another light. In the electronic medium they can be easily – and again post festum – assembled to comprehensive aggregates of knowledge. It might sound scandalous in the ears of conservatives, but even in digitally based scholarship Clay Shirky might be right with the idea just quoted: »publish first, filter later.«

The significance of what I have tried to shortly mention in this part of my talk cannot be overrated, because it will also entail deep cultural consequences. For the sake of clarification I refer you to Chris Anderson's important study on the »Long Tail«, a popular paraphrase of the Pareto-principle. (The long tail. Why the future of business is selling less of more, 2006) Few actors concentrate a huge mass on themselves. For the bookmarket, this means that there is a small number of books with high sellings and a huge number of them with low ones.

In the analogous economy we have to follow the rule for concentrating on the blockbusters, because the long tail – the millions of books available but mostly unknown – is logistically unmanageable if your bookstore is smaller than a Walmart-storehouse. And this is not just an economic constraint, but it also ensures mainstream-culture, because what is extraordinary, strange, maybe deviant, is hidden in the long tail where in traditional culture it tends to be forgotten – or not even published. But maybe not in the digital culture.

Referring this to scholarship: All that remains necessarily unpublished in the prior reviewing procedure, now has a chance. And this is not so far fetched as in scientific print-culture there is heavy criticism of the fact that much that is innovative falls through the cracks. This is certainly true for art history where what is established normally beats what is original. The problem of unmanageable masses would also be approached by such an innovative procedure. What is really nonsense or insignificant disappears in the depths of computer storage. But it exists and can – when its time will eventually have come – reemerge from there and leave behind the long tail.

Conclusion

We seem to have gone a long way from social tagging to scientific publications. But the structural parallels are obvious. In both cases the production of knowledge is assigned to a wider audience, although an audience eventually restricted in the most different and in each case pertinent way. And in both cases the quality management is assured by a downstream filtering process, whose nature is dependent on the sophistication of the software. I wouldn't regard this as a danger, but uphold its advantages.