Project no. 34721

TAGora

Semiotic Dynamics in Online Social Communities

http://www.tagora-project.eu

Sixth Framework Programme (FP6)

Future and Emerging Technologies of the Information Society Technologies (IST-FET Priority)

Periodic Activity Report

Period covered: from 01/06/2008 to 31/08/2009
Start date of project: June 1st, 2006
Due date of deliverable: October 15th, 2009
Distribution: Public

Date of preparation: 15/09/2009
Duration: 39 months
Actual submission date: September 15th, 2009
Status: Final

Project coordinator: Vittorio Loreto
Project coordinator organisation name: Sapienza Università di Roma
Lead contractor for this deliverable: Sapienza Università di Roma

Contents

Publishable executive summary

The vision

TAGora is a project sponsored by the Future and Emerging Technologies program of the European Community (IST-034721) focussing on the semiotic dynamics of online social communities. The widespread diffusion of access to the Internet is making possible new modalities of interaction between Web users and the information available online. The new vision of the Web regards users not only as producers or consumers of information, but also as architects of the information on the Web, which gets shaped according to criteria closely related to the meaning of information, the semantics of human agents. In this perspective the Web is becoming an infrastructure for “social computing”, that is, it allows to coordinate the cognitive abilities of human agents in online communities, and steer the collective user activity towards predefined goals.

Figure 1: Logo of TAGora project. Web site at www.tagora-project.eu

An approach to information management that has become wildly popular during 2005 (in a matter of a few months), is *collaborative tagging*. The central idea is that users interested in organizing and sharing a certain kind of resources (digital photographs, web pages, academic papers, and so on), use a web application to associate free-form keywords – called “tags” – with the content they’re interested in. Such associations are personal, but globally visible to the user community. At the system level the set of tags, though determined with no explicit coordination, evolves in time and leads towards patterns of terminology usage that are shared by the entire user community. Hence one observes the emergence of a loose categorization system – commonly referred to as *folksonomy* – that can be effectively used to navigate through a large and heterogeneous body of resources. Tags act as a sort of “semantic glue” bringing together resources and users in a time-dependent and truly complex architecture, providing an unexpected bottom-up realization of the semantic web vision originally proposed by Tim Berners-Lee.

Overall, the collaborative character underlying many Web 2.0 applications puts them, very naturally, in the spotlight of complex systems science, since the problem of linking the low-level scale of user behavior with the high-level scale of global applicative goals is a typical problem tackled by the science of complexity: understanding how an observed emergent structure arises from the activity and interaction of many globally uncoordinated agents. The large number of users involved, together with the fact that their activity is occurring on the Web, provide for the first time a unique opportunity to monitor the “microscopic” behavior of users and link it to the emergent properties of Web 2.0 applications (for example the global properties of a folksonomy) by using formal tools and conceptual frameworks from Statistical Physics. Understanding how the emergent properties of applications are linked to the behavior of their users is a challenging problem at the interface of several fields, from computer science and complex systems science, to cognitive science and information architecture. TAGora project aims at understanding and modeling information dynamics in online communities, providing a solid scientific foundation for the emerging field of “Web Science”.

Scientific and Technological Objectives

The project is articulated in four main areas whose activities are strongly intertwined. The initial phase of the project will deal with collecting actual data from existing, live systems and analyzing them with a variety of formal tools, eventually inferring models that are able to capture the essential features of the emergent dynamics, and explain how they might arise from the interactions of single agents. The inferred models of the emergent dynamics will be subsequently used to develop simulations that will allow the formulation of design strategies targeted at attaining a specific global behavior.

Emergent metadata The initial phase of the project will deal with collecting actual raw data from existing, live systems. By “raw data” we mean the emergent metadata that arise because of agent interactions in online social communities, as described in the introduction. Several online communities are readily accessible over the web: for a selected set of these systems, tools will be developed and deployed to harvest the relevant data, metadata and temporal dynamics, and to store the acquired information in a form amenable for data analysis.

Data analysis of emergent properties Examining quantitative aspects of folksonomy is a highly important area of research. Our objective is the set up of several protocols of data analysis to be performed on the raw data sets. A data analysis protocol is defined by: (1) indicating a specific quantity / observable / estimator suitable of a quantitative measure on the raw data sets; (2) acquiring the existing software tools, or developing new specific tools, needed to perform the measure; (3) extracting the relevant statistical information characterizing the analyzed data sets.

The aim of the data analysis is to identify and quantify emergent properties of the system in study, i.e. properties that can not be simply inferred from the behavior of the single agent. Beyond suggesting the collection of new or more refined raw data, the results of the data analysis will be used to

- ⋆ identify general features common to the different systems in study
- ⋆ characterize/discriminate the specific features of different systems in study
- ⋆ orient the modelling phase of the research project (see below)
- ⋆ providing benchmarks to test/improve existing systems or to suggest the creation of new more performing systems

Modeling and simulations

The objectives of this research area are twofold:

understanding complexity: develop models that captures the essence of the emergent dynamics and explain how it might arise from the interactions of single agents;

taming complexity: formulate design strategies that allow controlling the behavior of the system at the emergent level by suitably choosing the microscopic dynamics of the interacting agents

One of the most important goals is to construct, implement and study specific modeling schemes aiming at reproducing, predict and control the emergent properties seen in the semiotic dynamics orchestrated in on-line communities. We plan in particular a modeling activity at different scales. On the one hand it will be important to construct microscopic models of communicating agents performing language games without any central control. At a different scale we shall consider more coarse-grained probabilistic models. Several models will be proposed to address specific aspects/scales of folksonomy. The models will allow computer simulation aimed at measuring emergent features to be compared with the results of the data analysis activity. The simulations should give an insight in how users select tags, what kind of categories and category structures underlying the evolving system of tags, how categories and tags are related to the objects being tagged, etc. It will also give information on what kind of more global structures (such as the most frequent tags) can be provided to users to optimize their on-line community infrastructure. The models will require components for assigning or adopting tags, categorizing data, and collective dynamics. However the approach will be to keep the models as simple as possible, identifying the minimal ingredients responsible for the emergent properties. The minimal character of the models should make a more analytical mathematical study feasible.

A possible way to tackle the complexity of the systems is to individuate different time scales, which can be separated. For instance, we expect that the dynamics of the social network of the folksonomy could be different from the time scale of the dynamics of the resources and of the tags. In this case one can, as a first approximation, propose a model of tags and/or resource dynamics based on a given, slowly evolving social network topology. This kind of assumption should be tested and corroborated as much as possible with the observations coming from the real data analysis.

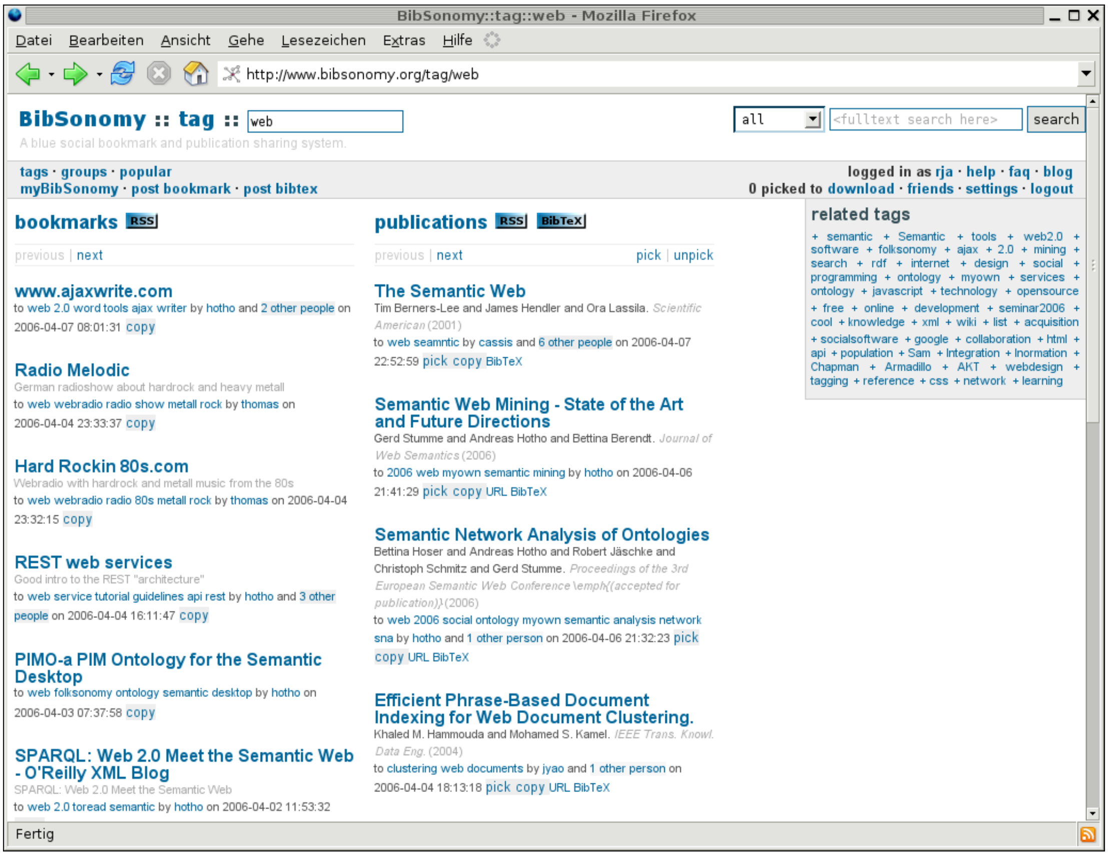
Feedback and control Finally, the output of all these activities has the potential to feed back into the data collection activity, specifically to the live social tagging system developed as part of, in order to experimentally verify the devised control strategies and demonstrate the technological advantage achieved by the present project.

Long-term applications

Collaborative tagging originated from the need to manage large collections of data. Tagging data is a means to describe, search, and retrieve objects in an intuitive way, which constitutes an important factor of its success. On the long term TAGora will provide experimental systems which are on the one hand intended to further improve navigation possibilities provided by tags, and on the other hand deliver data for the research work of the project. In order to have privileged and controllable data sources for the collaboration, TAGora planned to design and deploy systems - both online systems and actual demonstrations/experiments - for the specific purpose of data collection. The first objective involves building systems that add value to existing tagging sites. One possibility is to enrich navigation based on tags by adding data analysis. The combination of data features and tagging allows to overcome shortcomings of tag-based search, such as problems caused by synonymy, homonymy, missing tags, or spelling mistakes. The added value of original TAGora systems is important in order to attract users and thus fulfil our second objective: to serve as a valuable source for data delivery. Moreover the new applications will allow the Consortium to gain unimpeded access to the raw data and will ultimately provide an experimental “clean room” platform that will be employed to validate the understanding of metadata emergence, and to experiment innovative control strategies. During the first year TAGora deployed prototypical versions of applications.

BibSonomy

BibSonomy (`http://bibsonomy.org`, see Fig. 2), allows users to upload their bookmarks or bibliographic references and assign them arbitrary labels, denoted “tags”. Moreover users may share bookmarks and publication references. In general, social resource sharing systems all use the same kind of lightweight knowledge representation, called folksonomy.


BibSonomy::tag::web - Mozilla Firefox
Datei Bearbeiten Ansicht Gehe Lesezeichen Extras Hilfe
http://www.bibsonomy.org/tag/web
BibSonomy :: tag :: web
all
<fulltext search here>
search
tags · groups · popular
myBibSonomy · post bookmark · post bibtex
logged in as rja · help · faq · blog
0 picked to download · friends · settings · logout
bookmarks RSS
publications RSS BibTeX
previous | next
pick | unpick
related tags
Fertig

A blue social bookmark and publication sharing system.

bookmarks

www.ajaxwrite.com
to web 2.0 word tools ajax writer by hotho and 2 other people on 2006-04-07 08:01:31 copy

Radio Melodic
German radioshow about hardrock and heavy metall
to web webradio radio show metall rock by thomas on 2006-04-04 23:33:37 copy

Hard Rockin 80s.com
Webradio with hardrock and metall music from the 80s
to web webradio radio 80s metall rock by thomas on 2006-04-04 23:32:15 copy

REST web services
Good intro to the REST "architecture"
to web service tutorial guidelines api rest by hotho and 3 other people on 2006-04-04 16:11:47 copy

PIMO-a PIM Ontology for the Semantic Desktop
to web folksonomy ontology semantic desktop by hotho on 2006-04-03 07:37:58 copy

SPARQL: Web 2.0 Meet the Semantic Web - O'Reilly XML Blog
SPARQL: Web 2.0 Meet the Semantic Web
to web 2.0 toread semantic by hotho on 2006-04-02 11:53:32

publications

The Semantic Web
Tim Berners-Lee and James Hendler and Ora Lassila. *Scientific American* (2001)
to web seamntic by cassis and 6 other people on 2006-04-07 22:52:59 pick copy BibTeX

Semantic Web Mining - State of the Art and Future Directions
Gerd Stumme and Andreas Hotho and Bettina Berendt. *Journal of Web Semantics* (2006)
to 2006 web myown semantic mining by hotho on 2006-04-06 21:41:29 pick copy URL BibTeX

Semantic Network Analysis of Ontologies
Bettina Hoser and Andreas Hotho and Robert Jäschke and Christoph Schmitz and Gerd Stumme. *Proceedings of the 3rd European Semantic Web Conference \emph{(accepted for publication)}* (2006)
to web 2006 social ontology myown semantic analysis network sna by hotho and 1 other person on 2006-04-06 21:32:23 pick copy URL BibTeX

Efficient Phrase-Based Document Indexing for Web Document Clustering.
Khaled M. Hammouda and Mohamed S. Kamel. *IEEE Trans. Knowl. Data Eng.* (2004)
to clustering web documents by jyao and 1 other person on 2006-04-04 18:13:18 pick copy URL BibTeX

related tags

+ semantic + Semantic + tools + web2.0 + software + folksonomy + ajax + 2.0 + mining + search + rdf + internet + design + social + programming + ontology + myown + services + ontology + javascript + technology + opensource + free + online + development + seminar2006 + cool + knowledge + xml + wiki + list + acquisition + socialsoftware + google + collaboration + html + api + population + Sam + Integration + Inormation + Chapman + Armadillo + AKT + webdesign + tagging + reference + css + network + learning

Figure 2: Screenshot of BibSonomy

MyTag

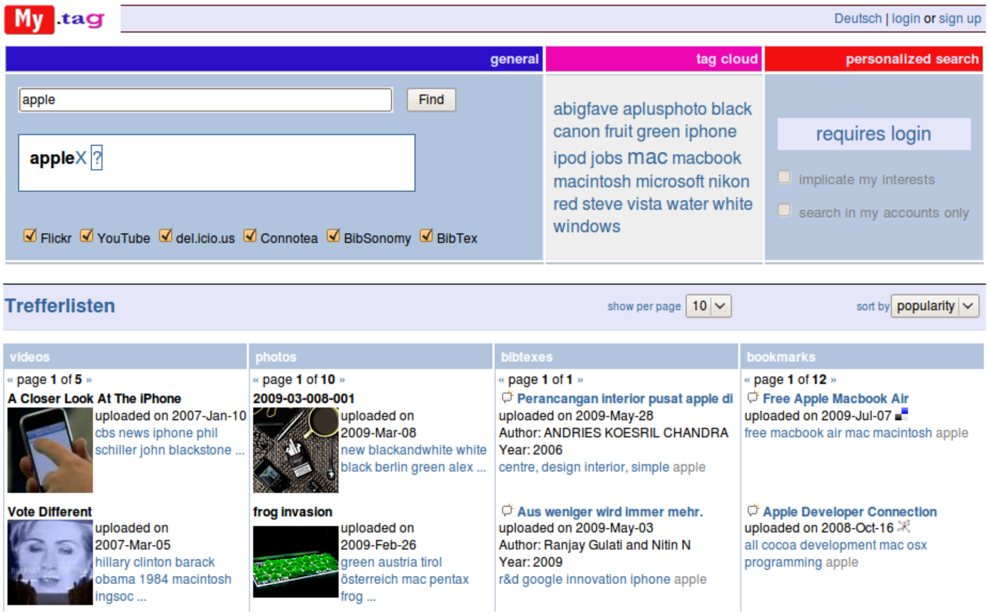
MyTag (`http://mytag.uni-koblenz.de`, see Fig. 3), aims at solving the limitations of current tagging platforms by searching different content types like photos, videos and social bookmarks from different sources in parallel. The search is transparently executed via the public APIs of the different tagging systems and the retrieved results are presented in separate columns for each content type. Besides, MyTag collects the search interests of registered users and offers them a personalized ranking of results. Additionally, an intelligent search assistant helps in disambiguating the current search terms by grounding them to possibly relevant articles found in DBPedia.

Ikoru

Ikoru (`demo.ikoru.net`, see Fig. 4) is a prototypical system that unifies browsing by tags, visual and audio features. This combination allows an intuitive exploration of databases and helps to overcome shortcomings of solely tag-based systems. In contrast to traditional image retrieval approaches, Ikoru employs user tags, complemented by image and music data analysis and classification.

NoiseTube

NoiseTube (http://noisetube.net, see figure 5), is the continuation and consolidation of the Zexe.net project. Like Zexe.net, the focus remains on supporting offline communities with new applications of collaborative tagging in real world situations, in particular for sustainability-related issues, by using mobile phones. Concretely, the NoiseTube platform enables a new participatory approach to monitor noise pollution by turning mobile phones into noise sensors and making intensive use


My.tag
Deutsch | login or sign up
general
tag cloud
personalized search
apple
Find
appleX ?
☑ Flickr ☑ YouTube ☑ del.icio.us ☑ Connotea ☑ BibSonomy ☑ BibTex
abigfave aplusphoto black canon fruit green iphone ipod jobs mac macbook macintosh microsoft nikon red steve vista water white windows
requires login
□ implicate my interests
□ search in my accounts only
Trefferlisten
show per page 10
sort by popularity
videos
photos
bibtexes
bookmarks

« page 1 of 5 »
A Closer Look At The iPhone
uploaded on 2007-Jan-10
cbs news iphone phil schiller john blackstone ...

Vote Different
uploaded on 2007-Mar-05
hillary clinton barack obama 1984 macintosh ingsoc ...

« page 1 of 10 »
2009-03-008-001
uploaded on 2009-Mar-08
new blackandwhite white black berlin green alex ...

frog invasion
uploaded on 2009-Feb-26
green austria tirol österreich mac pentax frog ...

« page 1 of 1 »
Perancangan interior pusat apple di
uploaded on 2009-May-28
Author: ANDRIES KOESRIL CHANDRA
Year: 2006
centre, design interior, simple apple

Aus weniger wird immer mehr.
uploaded on 2009-May-03
Author: Ranjay Gulati and Nitin N
Year: 2009
r&d google innovation iphone apple

« page 1 of 12 »
Free Apple Macbook Air
uploaded on 2009-Jul-07
free macbook air mac macintosh apple

Apple Developer Connection
uploaded on 2008-Oct-16
all cocoa development mac osx programming apple

Figure 3: Screenshot of MyTag

of tagging. By using off-the-shelf smart phones, we empower the general public to measure, annotate, geo-localise and share their personal exposure to noise to inform the community and build collaborative, semantically annotated noise maps. The NoiseTube platform and the underlying approach has been presented and published at several conferences and workshops. Maisonneuve et al. (2009a,b,c) and received a great deal of attention from environmental agencies (e.g. Bruit-Parif in Paris, France and Awaaz Foundation in Mumbai, India). It was also covered in an article on a popular science website (Westly (2009)).

Results achieved so far

The main results achieved so far include:

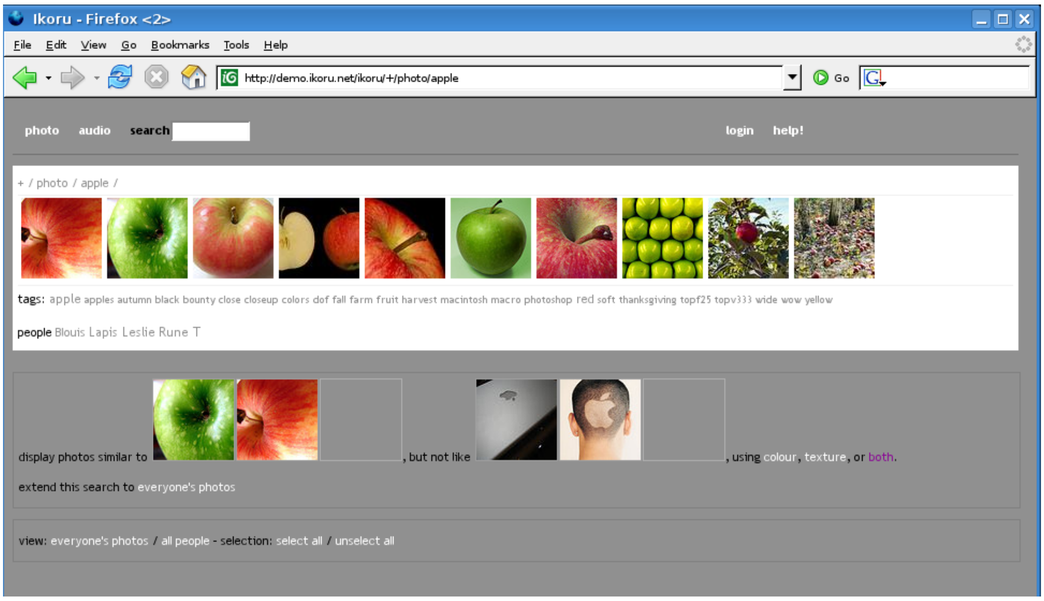
WP1 Consolidation of the applications to survive and increase the dissemination: new API to reuse Ikoru's features using Flickr, consolidation of zexe.net into NoiseTube;

WP1 extensive data collection from selected collaborative tagging systems (Full snapshot of del.icio.us and large scale snapshot for Flickr and Last.Fm);

WP1 acquisition of existing datasets from several social websites (IMDB, Netflix, Wikipedia);

WP2 realization of the application NoiseTube, a platform that enables a participatory approach to monitor and map noise pollution by turning mobile phones into noise sensors and making intensive use of collaborative tagging.

WP3 The designers of tagging systems have been defining and using several measures to detect and propose "related" tags. The notion of relatedness, however, has always been unclear. We investigated in a systematic fashion several aspects of similarity and relatedness in social bookmarking systems, using data from popular systems (Delicious) as well as from systems developed by the Consortium (Bibsonomy). We investigated tag relatedness by grounding graph-based notions of similarity computed on the folksonomy in formal representations of knowledge such as WordNet (for terms) or the Open Directory project (for URLs). This



http://demo.ikoru.net/ikoru/+/photo/apple

+ / photo / apple /

tags: apple apples autumn black bounty close closeup colors dof fall farm fruit harvest macintosh macro photoshop red soft thanksgiving topf25 topv333 wide wow yellow

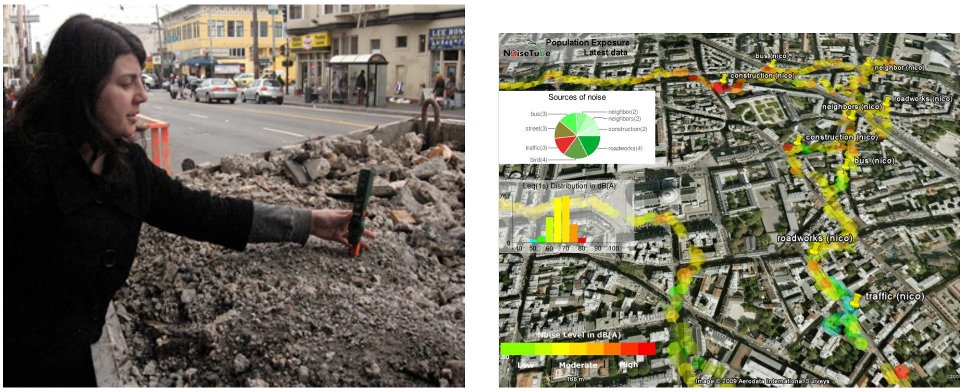
people Blouis Lapis Leslie Rune T

display photos similar to , but not like , using colour, texture, or both.

extend this search to everyone's photos

view: everyone's photos / all people - selection: select all / unselect all

Figure 4: A screenshot of Ikoru from http://demo.ikoru.net



(a) A NoiseTube participant measures and annotates her noise exposure using her mobile phone

(b) The visualisation of a noise exposure map in a district with a semantic layer to identify the sources of noise thanks to tagging features

Figure 5: Screenshot of NoiseTube

produced insights into the semantic character of tag similarity Cattuto et al. (2008a,b) and resource similarity Markines et al. (2009), and paved the door to control strategies based on suggesting tags or resources that bear a desired semantic relation with one another. Some of these strategies have been implemented in Bibsonomy as well as in a distributed social bookmarking system (GiveALink.org) developed by collaborators at Indiana University. Grounded notions of similarity in folksonomies were leveraged in WP4 to create unsupervised and semi-supervised spam-detection systems.

WP4 Detailed test of the Semantic Walker Model by considering networks with very different topologies in order to test the robustness of our approach with respect to the structure of the underlying network. The work has been published on the Proceedings of the National Academy of Sciences Cattuto et al. (2009).

WP4 Implementation of the Semantic Walker Model using a real Word Association graph a proxy for a latent shared semantic graph. We considered in particular the South Florida Free Association Norms database (`http://w3.usf.edu/FreeAssociation/`).

WP4 Implementation of innovative control strategies on existing applications, leveraging on theoretical analysis and models. In particular Bibsonomy was used to deploy recommendation schemes that leverage our new understanding of semantic similarity in tagging systems. The navigation interface of Bibsonomy now provides advanced tag recommendation, allowing the user to move from tag to tag in semantically-controlled directions (towards more general tags, towards similar/synonym tag). Similar users are also suggested by the system to foster the growth of social networks based on shared interests inferred from metadata. The recommendation framework was deployed together with a spam-detection framework, which is also based on novel spam detection techniques devised by the TAGora project. A logging framework for the user-interface events of Bibsonomy has been deployed to enable user studies aimed at evaluating the effects of the above-mentioned features on user behavior, navigation patterns, and information foraging.

WP5 Organization of the Hypertext 2009 conference, the 20th ACM conference on hypertext and hypermedia, held in Turin from June 29th to July 1st 2009. (`http://www.ht2009.org/index.php`). Ciro Cattuto was the general co-chair of the conference, together with Giancarlo Ruffo from the University of Torino. Andreas Hotho and Vittorio Loreto were the chairs of a track devoted to *People, Resources, and Annotations*. Hypertext 2009 was an excellent showcase of TAGora achievements, in terms of both paper presentations and technology demos.

WP5 Organization of a big TAGora-sponsored workshop on *Tagging Dynamics in Online Communities* in the framework of Hypertext 2009, June 29th, Turin, Italy. The workshop, open to all the other projects of the cluster, has been an excellent opportunity for all the TAGora team to showcase thei main achievements during the three years of the TAGora project.

WP5 Realization of the Live Social Semantics (LSS) experiment where the Semantic Web, the Social Web, and the Physical World come together to create a rich and integrated network of information. Live Social Semantics (LSS) is an application developed in collaboration with the SocioPatterns.org RFID project to disseminate TAGora technologies at the European Semantic Web Conference 2009. Because of its success, LSS was deployed again at Hypertext 2009. LSS was a great dissemination activity for the project, and we have ambitious plans for taking it much forward, to other bigger conferences, as well as to large corporate, business, and scientific events.

Consortium and contact details

The project is coordinated by Vittorio Loreto (Physics Dept., *Sapienza* Universià di Roma) and includes the following partners and node coordinators:

- Physics Department, *Sapienza* Università di Roma (PHYS-SAPIENZA), Italy, Vittorio Loreto
- Sony Computer Science Laboratory (SONY-CSL), France, Luc Steels
- University of Koblenz-Landau (UNI KO-LD), Koblenz, Germany, Steffen Staab
- University of KASSEL (UNIK), Kassel, Germany, Gerd Stumme
- University of Southampton (UNI-SOTON), Southampton, UK, Nigel Shadbolt

Please contact:
Vittorio Loreto, Physics Dept., *Sapienza* Università di Roma
Tel: +39 06 4991 3461

E-mail: vittorio.loreto@roma1.infn.it
For more information see: http://www.tagora-project.eu

Chapter 1

Project objectives and major achievements during the reporting period

1.1 Project objectives

A new paradigm has gained impact in large-scale information systems: Social Tagging. In applications like Flickr, Connotea, Citeulike, Delicious, etc. people no longer make passive use of online resources: they take on an active role and enrich resources with semantically meaningful information. Such information consists of terminology (or “tags”) freely associated by each user to resources and is shared with users of the online community. Despite its intrinsic anarchist nature, the dynamics of this terminology system spontaneously leads to patterns of terminology common to the whole community or to subgroups of it. Surprisingly, this emergent and evolving semiotic system provides a very efficient navigation system through a large, complex and heterogeneous sea of information.

Our project is aimed at giving a scientific foundation to these developments, so contributing to the growth of the new field of Semiotic Dynamics. Semiotic Dynamics studies how semiotic relations can originate, spread, and evolve over time in populations, by combining recent advances in linguistics and cognitive science with methodological and theoretical tools of complex systems and computer science.

The project is exploiting the unique opportunity offered by the availability of enormous amount of data. This goal will be achieved through:

(a) a systematic and rigorous gathering of data that will be made publicly available to the consortium and to the scientific community;

(b) designing and implementing innovative tools and procedures for data analysis and mining;

(c) constructing suitable modeling schemes which will be implemented in extensive numerical simulations.

We aim in this way at providing a virtuous feedback between data collection, analysis, modeling, simulations and (whenever possible) theoretical constructions, with the final goal to understand, predict and control the Semiotic Dynamics of on line social systems.

1.2 Objectives and main achievements of the reporting period

PHYS-SAPIENZA In the 3rd and last year, we went on performing the statistical analysis and modeling of the datasets so far collected by the consortium.

In particular:

- We studied the semantic assortativity in the social networks hosted by `Flickr` and observed a clear assortativity, larger than in suitable null model adopted for comparison. This semantic similarity, however, did not appear to develop during the community evolution, but seems rather the result of a common pre-existing background knowledge between users.
- We explicitly consider the dynamics of tagging, with further study on bursty tag streams in a comparative study over different folksonomies.
- We extended the random-walk based model of folksonomy generation with an extensive evaluation on different semantic network topologies. Correspondingly we start new statistical data analysis on the post structure.
- We introduced the directed network version of a folksonomy by using tags order in the post and defined the possible cosine-like similarity measures for such a system. By comparing the actual directed network with the directed reshuffled one, we found that the tags order in a post may contain a semantic signature.
- We consider the dynamics of social networks in folksonomy systems, comparing the dynamics of `Flickr` groups with the exposed network of user contacts.
- We studied the group membership statistics and in `Flickr`, uncovering non-trivial properties of the explicit social network connecting users in a same group.

SONY-CSL For the 3rd year, one of the main objectives was to focus on an action plan for the developed applications Ikoru and Zexe.net. (through its successor NoiseTube). The goal was to ensure their survival and re-usability outside the TAGora project. For Ikoru, in the second Periodic Activity Report SONY CSL declared its decision not to promote the website to a large audience. The reason for this decision was that maintaining an active multimedia Web site would require too many resources. Taking into account the reviewers comments, SONY CSL decided to provide a lightweight API for Ikoru to make the main feature such as the hybrid classifier mixing tagging and low-level features for images accessible to Flickr. Turning Ikoru into a web service with original features for Flickr increases its chance to attract a larger audience.

The Zexe.net platform supporting new tagging applications in real-world situations and for sustainability-related issues using mobile phones, was consolidated and extended to the case of noise pollution and was renamed to *NoiseTube*. NoiseTube is thus a platform which enables a participatory approach to monitor and map noise pollution by turning mobile phones into noise sensors and making intensive use of collaborative tagging. By using off-the-shelf smart phones, we empower the general public to measure, annotate, geo-localise and share their personal exposure to noise to inform the community and build collaborative, semantically annotated noise maps.

Concerning the WP3 and the music analysis, the first result addressed the issue of automatic prediction of tags from the acoustic analysis of audio signals. Many interesting scientific results were obtained and published. However, the results were shown to be globally insufficient for completely automatic predictors. They were, therefore, not implemented in the Ikoru framework. The other activity consisted in inventing a new paradigm for exploiting user tags, called Description-Based Design. In this context, tags are used as "control handles" to steer object generation mechanisms. A successful pilot study was conducted in the field of musical composition and published in the Computer Music Journal Pachet (2009).

Since 2008, SONY-CSL has been collaboration with photographer Armin Linke for his "Phenotypes/Limited Forms" installation which is built around the Ikoru platform. This installation allows each visitor to print his own photo album from a set of Armin Linke's photos and tag it. After more than a year and several expositions in the world, the "Phenotypes/Limited Forms" installation gathered around 190.000 tag assignments. In the 3rd year, after more than one year of data collection, its analyse was initiated. A cooperation was established with the UNIK team to use their BibSonomy system with the dataset.

UNI KO-LD

In the final year we decided to concentrate our development efforts on a single application. We choose MyTag since it is a flexible system which is also well suited for integrating and testing the various findings from the research in WP3 and WP4. Therefore, two main objectives were followed: (i) a general improvement of the MyTag system in terms of performance and extensibility and (ii) the integration of new features and enhancements like improved personalized ranking and tag disambiguation support.

In WP3, the analysis of the tagging data had made good progress in the second year especially with respect to categorizing tags. We planned to intensify the efforts in semantic inferencing and extend our approaches in the direction of classifying users, tags, and resources. One interesting aspect turned out to be the classification of landmark photos. Moreover some more large-scale evaluations have been done to verify the findings with bigger folksonomy data sets.

We had also observed that the sparseness of tagging data is hampering the search result quality in tagging system. One way to solve this problem would be the enrichment of the tagging data with missing information. We wanted to investigate how this could be done, e.g. based on user or tag co-occurrence.

Finally, the epistemic dynamic model in WP4 should be further extended, e.g. to integrate the simulation of complete tag postings, and be verified with different data sets. This model already proofed to be valuable and so the findings could be directly feed back into the analysis in WP3. This model is able to characterize a users tagging behaviour so the idea was whether this could be useful for identifying spammers due to their abnormal tagging activity.

UNIK Our major objective for the last year was to close the cycle from the applications to models and back to applications. To this end, we have continued our research in the following areas: measures for the semantic similarity of tags, tag and user recommendations, spam detection, and usage mining.

The functionality of BibSonomy was further enhanced, as described in Deliverable 2.5. In particular, the best of measures resp. algorithms have been deployed in BibSonomy, namely as display of related tags, similar tags, and related users. Furthermore, we implemented a logging framework and a spam framework, and set up the design of the recommender framework. The latter had to be implemented outside of TAGora, financed by a national follow-up project, due to the limited TAGora resources. The frameworks were the technical platform for the experiments performed in WP 4.

UNI-SOTON Much of our effort in the last year was focused on developing further, and deploying, various services. One such service is dedicated to Tag Filtering, which now any application developer can use to clean (or normalise) sets of tags to enable better integration and matching of tags and resources. This service can be of vital importance for any work that involves tag analysis, where such tags usually suffer from well known problems of synonymy, different format of the same term (e.g. blog, blogs, blogger, bloggins), compound terms, misspellings, etc. With our Tag Filtering services, anyone can submit a set of *raw* tags, and receive a set of *filtered* tags, where such terminological and syntactical variations are fixed by replacing all tags with their canonical form.

Another service we opened up is for Sense Matching. Same tags are often used for different meanings, which is another well known problem in the domain. This service can be called to find DPpedia resources and Wordnet synsets that may correspond to the meaning of a given tag, along with vector space models of tag senses to support tag disambiguations.

The third service that we extended and made accessible is for Profile Linking, which allows users to link their multiple online identifies and tagging activities. Another is the Profiling Building service, which uses the above services to suggest interests to users, distilled from their tagging activities across multiple folksonomies.

In this reporting period, we also spent effort on researching and developing a tag disambiguation algorithm, whose goal is to rank and select DBpedia entries representing the meaning of the analyzed tag. The algorithm uses the Sense Matching service above, to retrieve a vector representation of the tag context and of the candidate DBpedia entries using a common vo-cabulary, based on term frequency of the candidate DBpedia entries. These vectors are compared using a common cosine-based similarity measure, and the most similar candidate DBpedia entry is selected as the tag meaning.

In the Project Activity Report of last year, we described our plan to launch a web site where users can view how their distributed tagging activities across various folksonomies (e.g. Delicious, Flickr, and Last.fm) have been compared and merged into user profiles of interests. This plan was put in place following the recommendation given by the review committee in year two of TAGora, which was to gather user feedback to better evaluate some of the project results.

To this end we built a system, Live Social Semantics (LSS - Alani et al. (2009)), that combines several TAGora-developed technologies to process data from social media with a platform, developed by the SocioPatterns.org project, that mines real-world social contacts using RFID devices. The system was made available to the attendees of two conferences: the European Semantic Web Conference 2009, and ACM HyperText 2009. Conference attendees were able to see how their tags have been collected from multiple sources, and filtered, disambiguated, and merged, and on how their interests have been inferred from their tagging activities. Live Social Semantics was built as a large dissemination window for our technologies. Several hundreds of users were able to use and learn about our TAGora services and developments. More detail in D2.5 and D4.5.

1.3 Summary of recommendations from previous reviews and brief description of how they have been taken up by the consortium

Here following is the review report n. 2 after the second year of activity of the project.

REVIEW REPORT N. 2 (covering period 01/06/2007-31/05/2008):

1 OVERALL ASSESSMENT OF THE PROJECT

"The Project TAGora proceeds according to the Annex 1 of the Contract in terms of Work Packages and Deliverables that have been foreseen. In its second year, the project continues to perform at the forefront of the state-of-the-art in all areas of work, including data collection, application and tool development, analysis and modelling. However, compared to the expectations set by the performance in the first year of the project, the improvements shown are rather incremental. This relative lack of success is largely due to the difficulty in implementing the feedback loop between modelling and applications development, which has been already identified in the first report as a critical requirement for major breakthroughs.

The final year of the project will need to concentrate efforts to realize this link in the remaining time and at the some time implement steps to insure the successful preservation of the existing results. Realizing this dual goal will require reinforced efforts in coordination and management: a closer collaboration among project partners is required and the effort will need to be shifted from areas *where the project is overachieving (in particular, application development) toward realizing these goals.*

In summary, the project is still well managed, from a financial as well as from a scientific point of view, but it cannot be relaxed during the last year of the project; otherwise, the expectations raised will not be fulfilled".

2 PROJECT ACHIEVEMENTS and FUTURE PLANS

2.1 Work carried out in the previous reporting period:

Scientific and Technical work

Workpackage 1:

"WP1 was clearly ahead of schedule after the first year of the project. Datasets are still an important asset of the project, along with the tools to reproduce them. Deliverables D1.1 (data from selected folksonomy sites) and D1.3 (data from selected recommender systems), whose data of delivery was month 11 were completed. At that time, other deliverables that were expected later had been already advanced.

In the second year, the collection of data proceeded at a smaller scale, although according to the needs of the project. The deliverables expected at the end of second year is D1.2 and it is about date from bibliographic references sharing systems and from experimental tag-based navigation systems. According to this, data has been collected from different sites. The title of the delivered documentation induces some confusion since it is entitled, ”Data collection from bibliographic references sharing systems”, but only item 1.8 can be considered as a bibliographic one. The other sets of data correspond mainly to tagging systems whereas integrating IMDB and NetFlix Datasets is more likely to be considered in the recommender systems tasks, already completed 1 year before. Since task remaining to be done in this WP1 is the public delivery of the data, this WP can be considered on schedule.

Following the recommendations of the previous year, those data sets that did not pose legal challenges were made available to the public and for this purpose a web page has been created.

Unfortunately, the visitor is still required to visit the websites of the partners to access the actual data sets and perform various steps to get to the data. There is also a large heterogeneity in the data formats used. I would recommend the project to centralize the data, to reduce the number of formats used (preferably converting all data sets to single format such as RDF) and in addition to data sets publish also the tools that allows others to reproduce the data. This last point is crucial as the data will age faster than the tools required to reproduce it. Data sets that have been held back from publishing should also be disclosed before the end of the project.

The methods for data filtering and for mapping profiles across folksonomies should also be made available as applications or web services. In the reviewers, estimate these methods add significant value to the raw data".

Description of how the suggestions/recommendations regarding WP1 has been taken into account.

As described in last report, a dedicated page[1] was added to the project web site for the description and sharing of several of TAGora's valuable datasets. This site has been updated with new datasets that were collected in the final year of the project.

Some of the datasets that TAGora had were held back from publishing, such as Delicious and

[1] http://www.TAGora-project.eu/data/

Flickr data. These datasets are now available for download by the public from the Data page of TAGora.

Currently the site contains 8 datasets, including large collections of Delicious, Flickr, Bibsonomy, and LastFM data. Following the recommendation of the reviewers, these datasets are now also available in RDF.

With regard to publishing the tools that the project used to crawl this data, we have published scripts for collecting data from Delicious and Flickr. These scripts can be used by anyone who is interested in collecting his/her tagging data from those sites. Note that such scripts can quickly run out of date and become unsuitable for collecting data from the relevant website. This is due to the frequent change in the publishing style or text formatting of those sites, which quickly renders such tools obsolete. The tools published on the TAGora site will be updated whenever possible to ensure their continuity.

In addition to publishing the datasets and the data gathering scripts above, we have also developed and published services that, as part of their functionality, collect tagging and other data from various folksonomy sites, including Delicious, Flickr, and LastFM. Such data crawling is now embedded into the services described in section 1.2, which creates and links users profiles and generates lists of their interest in the form of DBpedia URIs.

Workpackage 2:

"WP2 is about applications. After finishing the first versions of the systems for sharing bibliographic data and for tag-based navigation systems for images and music, and Interim Report is expected. Further development of these systems has been achieved.

The work on applications is still over-achieving compared to the goals set forth in the Technical Annex. This problem was already noted in the previous report where the reviewers suggested shifting effort from extending features to implementing and experimenting with control mechanisms devised in WP4. Despite this recommendation the project has started developing a new application (MyTag).

The growing number of users of BibSonomy shows clearly the acceptance of the community. The facility for searching and the integration with other products has contributed to its success.

TAGster is a tagging system for sharing multimedia data and several improvements have been made along the second year. In any case, it seems difficult to reach a critical mass for an efficient use. The Tagster application is lagging behind in development and it remains unclear whether the project partners will be able to study tagging behaviour in a P2P environment at sufficient scale within the lifetime of project, and whether this investigation is worthwhile (given the lack of P2P systems that use tagging). Unless early results seem particularly promising, the effort to be spent on Tagster should be diverted to other tasks.

Looking at the applications as a whole, some of them are not integrated in the main body and the set looks like a collection of unrelated pieces.

The project needs to make a plan of actions as to what results of this Workpackage can survive the lifetime of the project and in what form. To the reviewers it seems that only Bibsonomy has gathered enough users to continue as it is. (Among others, Bibsonomy could be immediately useful for other projects in the FET cluster.) All or parts of the remaining applications could be potentially made available as open source where IPR allows. Key functionality could also be turned into lightweight APIs or Web Services so as to serve as a basis of other applications".

Description of how the suggestions/recommendations regarding WP2 has been taken into account.

During the third year of the project, applications developed within TAGora became crucial means to disseminate the research that is being made. These applications become the embodiment of the research itself, by reflecting the project's developments and new strategies.

BibSonomy, developed by the University of Kassel, has aimed to increase its user base even further and provide new services and features coming directly from the consortium's research. The system will remain live after the end of the project, financed by follow-up projects.

MyTag, the tag based multi-source search engine, which was developed by UNI-KO-LD, will be taken forward in a new application oriented EU project with UNI-SOTON. In this new project, we will use the ideas collected in MyTag for the purpose of eParticipation and eGovernment.

The work on Tagster has been discontinued in the third year due to the mentioned technical issues and problems to aquire a critical mass of users. However, Tagster has been published as open source at https://launchpad.net/tagster.

Concerning the Ikoru platform, developed by the SONY-CSL team, the source code was made available under an open source license (see http://ikoru.sourceforge.net), the web site will be kept online and Ikoru will continue to be used as part of the "Phenotypes/Limited Forms" installation, and SONY-CSL has developed a light-weight REST API. See deliverable 2.4 for further information.

The Zexe.net system, developed by the SONY-CSL team, will live on in its successor NoiseTube (http://noisetube.net). Like Zexe.net, the focus remains on supporting offline communities with new applications of collaborative tagging in real world situations, in particular for sustainability-related issues, by using mobile phones. Concretely, the NoiseTube platform enables a new participatory approach to monitor noise pollution by turning mobile phones into noise sensors and making intensive use of tagging. Development work on the NoiseTube platform will continue after the TAGora project and test deployments are currently being planned in cooperation with environmental agencies such as the Awaaz Foundation in Mumbai, India.

The Live Social Semantics system, developed by the University of Southampton and PHYS-SAPIENZA (through ISI Turin) on top of an RFID sensing platform developed by the SocioPatterns.org project, illustrates the possibilities of utilising various TAGora technologies for the analysis of the social behaviour of conference participants. It was deployed at two major conferences in 2009. We have already opened up most of the services behind this application, by making them accessible with HTTP queries and SPARQL. See Deliverables 2.5 and 4.5 for further details. We are currently seeking further funding to sustain this application way beyond TAGora.

Workpackage 3:

"WP3 concerns "Data analysis of emergent system properties". The analysis of emergent properties progressed steadily in the second year, although most of the improvements to existing models were incremental. It has also become increasingly clear that many of the methods produce results at the macro level (relating to distributions and rates of change) that seem difficult to translate into control methods at the micro level (see also WP4).

After having extracted emergent metadata statistics the tools required for the analysis were partially developed during the first year of the project. In particular, methods to identify communities in large folksonomies have been proposed. In any case, the documentation presented is too short and difficult to understand. The same definition of community is unclear since different definitions are proposed in the literature. Looking at the publications and contributions of the consortium about this issue, the progress carried out is considerable. Finally, it is worth mentioning that BibSonomy as an application already shows some of the results.

Finally, some consideration is due to the effort in linking cross-folksonomy networks. The portal MyTag integrates photos from Flickr, videos from YouTube, and bookmarks from del.icio.us through simple searches.

As mentioned for WP1, some of the methods developed in this workpackage are likely to be interesting for outside researchers, including data cleaning, cross-folksonomy mapping and semantic inference. These methods could be made available as services through the project website".

Description of how the suggestions/recommendations regarding WP3 has been taken into account by:

The theoretical analysis performed in the third year has developed analytical methods that could be converted in an applicative framework. For example, the detection of typical temporal patterns of the vocabulary length associated to a resource in a folksonomy as a function of the resource age can be used to discriminate between regular and spam contribution in a folksonomy.

In the work on enriching the vector space model we considered, besides tag distributions, some more specific aspects, namely resource-based and social tag contexts, i.e., which resources have a particular tag assigned and which users have used a particular tag. The integration of the contexts allowed us to include additional semantic information at a much finer granularity than with the macroscopic information derived from the tag distributions.

Furthermore, we used the Epistemic Model for designing model-based features for detecting spam in tagging systems. The assumption is that the model simulates an average non-spammer and that spammers show a different behaviour than the simulated behaviour. The features are thus based on comparing the real behaviour of users with the simulated behaviour. If the users show larger deviations from the simulated behaviour, they are likely spammers. The preliminary results reported in D4.6 show that the model-based features are able to identify non-spammers but also that the features have to be further improved in order to detect a larger percentage of the spammers.

Some of the services for cross-folksonomy analysis and integration have been made available as web services. This includes the *tag filtering service*[2] which can be sent a set of raw tags (e.g., an entire tag cloud), which will be processed and filtered by the service, and a clean set of tags will be returned back. The *sense matching service*[3] is a linked data enabled service endpoint that provides extensive metadata about tags and their possible senses. The other service we have opened up *RDF Builder*[4]. With this services, users can submit their account names for Flickr, Delicious, or lastFM, and receive back an RDF representation of all their tagging activities and any social ties they have with other members in those folksonomies. These three services formed some of the pillars of our cross-folksonomy work, and by opening access to them we hope to share that benefit with the community.

Finally, the research performed in TAGora has led to some implementations, eg, in BibSonomy. In particular, we have selected among the various similarity measures that we studied in the second project year the best ones for displaying related tags, similar tags, and related users. The research about spam detection and tag recommendations has found its way into BibSonomy as spam resp. recommendation framework. To further analyse the effects of these features on the user behavior, we have set up a logging framework.

Workpackage 4:

"WP4 is about "Modelling and simulation". The first year was devoted to review both theoretical tools for modellng and analyzing collaborative social tagging systems and existing recommendation strategies and systems. D4.2 is an Interim Report on models and simulations published at the end of the second year. The consortium is giving a quantitative overview of tagging systems by con-

[2] http://tagora.ecs.soton.ac.uk/tsr/tag_filtering.html

[3] http://tagora.ecs.soton.ac.uk/tsr/sense_matching.html

[4] http://tagora.ecs.soton.ac.uk/tsr/rdf_builder.html

structing models that explain folksonomy data. For instance from the stream view, the growth of the dictionary, the frequency rank distribution (analogue to Zipf's law) or the tag-tag correlations are studied; and from the network perspective, the co-occurrence network is observed, where nodes are tags connected if they are used in the same post.

Modelling activities showed solid improvements in the second year, allowing closer approximations of real tag co-occurrence graphs and tag streams. However, control strategies for implementing these results are still largely lacking. The lack of feedback from applications also means that the exploration of possible models for folksonomies and their dynamic behaviour is unguided. Some of the questions (such as the role of imitation vs. background knowledge) cannot be answered without explicit experiments. As experimentation through applications proves to be harder than expected an alternative would be to perform targeted user studies in a lab environment (e.g. comparing the behaviour of users exposed to previous tagging results vs. users who don't have access to such information.)

It is also presented a model of co-occurrence network where posts correspond to random walks of the user on a "semantic space". The results are robust under different choices of the parameters whenever the reference network is a Watts-Strogatz one but it is not clear at all if this robustness is preserved if other model networks are considered.

The reviewers value the initiative for the ECML/PKDD challenge on recommendation and spam detection in folksonomies. Even if the organizers themselves cannot directly participate, this challenge will allow validating project results against the state-of-the-art".

Description of how the suggestions/recommendations regarding WP4 has been taken into account by:

We payed a great attention to the remark about the difficulty to translate theoretical or modeling results into control methods at micro level. We stronly aimed at implementing innovative control strategies on existing applications, leveraging on theoretical analysis and models. In particular Bibsonomy was used to deploy recommendation schemes that leverage our new understanding of semantic similarity in tagging systems. The navigation interface of Bibsonomy now provides advanced tag recommendation, allowing the user to move from tag to tag in semantically-controlled directions (towards more general tags, towards similar/synonym tag). Similar users are also suggested by the system to foster the growth of social networks based on shared interests inferred from metadata. The recommendation framework was deployed together with a spam-detection framework, which is also based on novel spam detection techniques devised by the TAGora project. A logging framework for the user-interface events of Bibsonomy has been deployed to enable user studies aimed at evaluating the effects of the above-mentioned features on user behavior, navigation patterns, and information foraging.

As for the Semantic Walker Model we have considered networks with very different topologies in order to test the robustness of our approach with respect to the structure of the underlying network. In particular we considered Watts-Strogatz network model, a random scale-free network obtained from the uncorrelated configuration model, an homogeneous Erdős-Rényi random graph, and a model of strongly clustered scale-free network. The work has been published on a prestigious journal (see Cattuto et al. (2009)). In particular the robustness analysis are provided as a supplementary information joined to the paper. In addition to further test the robustness of the Semantic Walker Model we considered the South Florida Free Association Norms database (`http://w3.usf.edu/FreeAssociation/`) as a proxy for a latent shared semantic graph. On this database we first performed, in collaboration with Alain Barrat from CNRS, Marseille, a statistical analysis of the word association graph (WAG) (directed and weighted) in order to identify the so-called Strongly Connected Component (SCC) (i.e., the set of all the words among which all the possible pairs are connected by at least a possible path). We then performed the random walks experiment using the word association graph as a latent semantic graph in order to check the robustness of the results against a realistic implementation of the latent semantic graph. Notice the the word association graph is directed and weighted. As already done with artificial graphs we compared the statistical properties of the co-occurrence graph constructed with random walk procedures performed on the word association graph with those of real co-occurrence networks of del.icio.us. Also in this case a strong similarity between the properties of the artificial and real co-occurrence graphs emerge. More details are found in the corresponding section of this activity report.

Workpackage 5:

About "Dissemination and exploitation". "Basically referred to D5.4 which is the portal of the project. It keeps the basic information, well presented, easy to navigate".

Description of how the suggestions/recommendations regarding WP5 has been taken into account.

We kept paying a strong attention to the issues related to the dissemination of our results during the third year. The Final plan for using and disseminating the knowledge reports about our activites in this direction.

Resources

"The resources used for the work performed have been appropriate and allocated according to plan except for the following deviations: a. PHYS-SAPIENZA reported problems related to an offsetting procedure which if unresolved may affect the actual research effort they will be able to spend on the project in the next year. b. UNIK reported the leaving of a researcher and subsequent hiring difficulties. These deviations can be corrected in the coming year".

Description of how the suggestions/recommendations regarding the "Resources" has been taken into account by:

PHYS-SAPIENZA: The issue of the offsetting procedure is still unsolved due to the slow burocratic procedures among different faculties and departmemts of Sapienza University of Rome. The charge of the offsetting procedure has been kept entirely on the PHYS-SAPIENZA team with the idea that a solution will come only after the end of the project. During the last year the Physics Dept. anticipated the needed budget to ensure the normal research activity of the project.

UNIK: We have employed 1.5 PhD students and 0.5 technicians plus some students, so that both resources and workload are in balance at the end of the project.

3 CONSORTIUM PARTNERSHIP

"As in the previous period, there is a balanced workload within the project and each partner brings important contributions to the consortium. However, the reporting period also shows that cooperation could be improved. The fragmented application development is a display of each partner working individually and the result is a duplication of efforts. While scientific exchanges have taken place within the consortium, the number of joint publications is low".

Since 2008, SONY-CSL has been collaborating with photographer Armin Linke for his "Phenotypes/Limited Forms" installation which is built around the Ikoru platform. This installation allows each visitor to print his own photo album from a set of Armin Linke's photos and tag it. After more than a year of data collection and several expositions in the world, the "Phenotypes/Limited Forms" installation gathered around 190.000 tag assignments. In the 3rd year the analysis was initiated in cooperation with the UNIK team.

To improve the usability of the MyTag application, UNI KO-LD an UNI-SOTON have joint their effort in the integration of the tag sense disambiguation service as a basis for new features in Mytag. This enabled MyTAG to filter, disambiguate, and recommend tags and to their users. For further evaluations of the developed tagging models, a closer contact has been establised with PHYS-SAPIENZA to pursue a closer collaborations on these issues.

UNI-SOTON and UNI KO-LD have also started a collaboration during the development and launch of the Live Social Semantics experiment at ESWC, where Bibsonomy was integrated into the site to provide users with access to the conference paper collection inside Bibsonomy, and to encourage them to annotate them with tags. We plan to extend the Live Social Semantics system to allow users to enter their Bibsonomy account name so that their Bibsonomy tags and groups can be taken into account.

In addition to the above, UNI-SOTON has also collaborated with PHYS-SAPIENZA and UNIK to exploit state-of-the-art Information Retrieval methods for finding associations and dependencies between tags, capturing and representing differences in tagging behavior and vocabulary of various folksonomies, with the overall aim to better understand the semantics of tags and the tagging process [Benz et al. (2008)].

UNIK and PHYS-SAPIENZA, in turn, have continued their cooperation on evaluating measures for semantic tag similarities. The joint effort has focused on the research of semantic relations between the tags forming a single post. A modified tag co-occurrence network has been studied, where edges connect tags subsequent tags occurring in the same post, unveiling non-trivial features of users' activity leading to possible application in methods for tag recommendation. Such modified, directed network, is much sparser and easy to study numerically with respect to usual tag co-occurrence network, whose it maintains the basic semantic properties. Therefore, it turns to be a usable and meaningful scientific object for semantical analysis of folksonomies.

Finally, the seminar on Social Web Communities[5] in September 2008, organized by the UNIK team in the prestigious Leibniz Center for Informatics Schloss Dagstuhl was a focal point for researchers both within but also across the boundaries of the TAGora project. The objective of the seminar was to provide theoretical foundations for upcoming Web 2.0 applications and to investigate further applications that go beyond bookmark- and file-sharing. The main research question can be summarized as follows: how will current and emerging resource sharing systems support users to leverage more knowledge and power from the information they share on Web 2.0 applications? Research areas like Semantic Web, Machine Learning, Information Retrieval, Information Extraction, Social Network Analysis, Natural Language Processing, Library and Information Sciences, and Hypermedia Systems have been working for a while on these questions. In the workshop, researchers from these areas came together to assess the state of the art and to set up a road map describing the next steps towards the next generation of social software.

4 PROJECT MANAGEMENT AND CO-ORDINATION

"The management of the project continues to deliver excellent work in keeping the project within the financial bounds. Administrative tasks are also handled well except an overdue delivery of the PMR. The reviewers would recommend a change in the composition of deliverables which have been overly reduced, e.g. some deliverables for this period included figures for which the explanation appeared only in the papers. Presentations at the review meeting should also be numbered according

[5] http://www.dagstuhl.de/de/programm/kalender/semhp/?semnr=08391

to deliverables to make it easier for the reviews to follow the organization of the material.

Coordination activities within the project will need to be increased in the coming period to deliver the expected results by the end of the project (see above). It is also not too late for small-scale intra-cluster collaborations, and in particular the final results of the project could be delivered in coordination with other projects of the same cluster in order to reinforce the overall image and provide compelling arguments for further exploration of the scientific area.

Project management may apply for a three months extension to finalize the dissemination of the project results and align the planning with the remaining projects in the cluster.

A final Workshop, open to other related projects (STREPS in the proactive initiative on Complex Systems), and/or a book with the main contributions from the participants should be welcome".

We took into account the recommendations of the reviewers trying to keep all the deliverables more self-contained and self-consistent (as required with the second report) while keeping the overlla lengths moderate (as required with the first report). As for the structure of the review meeting we'll pay special care ti number in the correct way the different contributions in order to facilitate the evaluation of the presented material in the framework of what is expected.

As for the coordination activities to stimulate small-scale intra-cluster collaborations we have to admit that, despite the attempts, the differences in the themes on which the different projects of the cluster are focused on made very difficult the implementation of actual and concrete collaborations. From this point of view the three month extension we successfully applied for makes the timeline of our project on line with the those of the other projects of cluster, so that the review meeting could be jointly organized.

As for final actions to take we organized three main events: (i) the Hypertext 2009 conference, the 20th ACM conference on hypertext and hypermedia, held in Turin from June 29th to July 1st 2009. (`http://www.ht2009.org/`). Ciro Cattuto was the general co-chair of the conference and Andreas Hotho and Vittorio Loreto were the chairs of a thematic section devoted to *People, Resources, and Annotations*. (ii) In the framework of Hypertex 2009 we organized a big TAGora-sponsored workshop on *Tagging Dynamics in Online Communities* in the framework of Hypertext 2009, June 29th, Turin, Italy. The workshop, open to all the other projects of the cluster, has been an excellent opportunity for all the TAGora team to showcase thei main achievements during the three years of the TAGora project. (iii) Finally PHYS-SAPIENZA and UNI-SOTON, in collaboration with the ISI Foundation and with the Sociopatterns.org project, organized an ambitious social experiment where the Semantic Web, the Social Web, and the Physical World come together to create a rich and integrated network of information. Live Social Semantics (LSS) was extensivelt used to disseminate TAGora technologies at the European Semantic Web Conference ESWC2009. Because of its success we deployed LSS again at HyperText 2009. LSS was a great dissemination activity for the project, and we have ambitious plans for taking it much forward, to other bigger conferences, as well as to large corporate, business, and scientific events.

5 USE AND DISSEMINATION OF KNOWLEDGE

The dissemination activities of the project are in line with expectations. The scientific output is impressive although the number of joint publications is low. Dissemination through artistic installations and targeted communities (by Sony CSL) and dissemination through popular science media (by PHYS-SAPIENZA) are interesting ways of making the project results tangible for a wide audience.

Beyond these continuous activities the project will need to prepare a plan of action for the final use and dissemination of project results. Exploitable results include the data sets and tools for collecting, cleaning and correlating data, applications in source and binary forms, their components and user *communities, algorithms and tools for analysis and modelling and other results. As stated above, a final Workshop and/or edition of a book are convenient.*

Description of how the suggestions/recommendations concerning dissemination activities have been taken into account.

The action plan to disseminate the results of TAGora the projects has been achieved to match the recommendations of the reviewers.

Dissemination of datasets - WP1. A section “data” was created in the TAGora portal describing and publishing all the datasets. To facilitate their distributions we decided to publish the complex datasets in RDF format, and following this way the proposal of the reviewers. Some datasets underwent post processing with a clean up of recording artifacts, wrong or defective recorded data entries. They are accessible via a direct link, web APIs, or by sending an email to the owner due to licence agreement requirement. Furthermore data gathering scripts for Flickr and Delicious were updated and made public on the project site. These scripts enable anyone interested in collecting his/her data to retrieve an RDF representation of their tagging information, based on the TAGora Tagging Ontology.

Dissemination of applications - WP2. A section “products” was created in the TAGora portal describing the applications developed during the project such as BibSonomy, MyTag, Tagster, Ikoru, Zexe.net and NoiseTube. For the applications Ikoru and Tagster which have been made open source, the url of the source code or the svn repository is provided on the website, so the community can reuse and extend them. Additionally MyTag, Ikoru and NoiseTube also provide a lightweight APIs so their implemented features can be reused as a part of third party web services.

Dissemination of data analysis/modelling tools - WP3/WP4. Tools for tag assignment simulation and analysis of folksonomys are made available for other research community in the section "‘data analysis tool"’. The tool *Epistemic dynamical tagging model* is a generative tagging simulator by the UNI KO-LD team reproducing characteristic properties of tag streams. It is distributed as a downloadable package with a user manual. The tool *Net* to either analyse and generate random graphs.

SONY-CSL contribution for dissemination. The Sony CSL Open House, a bi-annual public symposium which was initially planned to take place in Paris in 2008, was replaced by an event in Tokyo for the occasion of the 20th anniversary of Sony CSL. Therefore an open, public presentation of the work done within the context of TAGora (Ikoru, Zexe.net and most importantly NoiseTube) was postponed until the Parisian Open House event now being planned for October 2009. At this event visitors will also be given the opportunity to experience the NoiseTube system first hand by making noise measurements and taggings in the street and seeing the result of their work on an interactive map afterwards.

The artistic installation of Armin Linke, titled "Phenotypes/Limited Forms", which is linked to the Ikoru system, continued to attract visitors at exhibitions during the 3rd year or TAGora. Currently it is still on display at an exhibition in the Museum of Contemporary Art in Siegen, Germany, which will remain open until 20 September, 2009.

PHYS-SAPIENZA As already mentioned above as final dissemination we organized three main events: (i) the Hypertext 2009 conference, (ii) the TAGora-sponsored workshop on *Tagging Dynamics in Online Communities* in the framework of Hypertext 2009, and (iii) the Live Social Semantics (LSS) experiment available to the attendees of two conferences: ESWC 2009, and Hypertext 2009. Users were able to see how their tags have been collected from multiple sources, and filtered, disambiguated, and merged, and on how their interests have been inferred from their tagging activities. Live Social Semantics was built as a large dissemination window for our technologies. Several hundreds of users were able to use and learn about our TAGora services and developments. More detail in D2.5 and D4.5.

6 CONCLUSION and SUMMARY of RECOMMENDATIONS

"Good to excellent project (The project has fully achieved its objectives and technical goals for the period and has even exceeded expectations)

Recommendation

The project should continue without modifications

Comments should include:

There should be two main targets of the project in the last year.

First, the integration of the project results and the implementation of the feedback between theoretical work and system development would need to continue with reinforced effort.

Second, with the end of the project in sight, the project members will need to ask themselves what the final outcomes of TAGORA should be and what compelling arguments can be made for continued funding of this research area.

We expect the consortium to deliver and execute a clear plan of actions that could ensure the survival of all results achieved during the project. This includes the measures for complete disclosure of the data and related tools in a way that it can reach the broadest range of users, the plan for the exploitation of applications so that they continue attract users or to be reused as part of other applications, and the dissemination of scientific results for the widest possible audience. This last part could be realized by a final workshop or a book.

In order to realize these goals, the project may divert effort from parts that already over-achieving, in particular application development".

PHYS-SAPIENZA The theoretical work developed in the third year open the path to possible technological applications. In particular, the detection of “significant tags” in a folksonomy, as described in task 3.1, can be applied in a tag recommendation system, in order to establish a suitable number of suggested tags to be submitted to a users according to a quantitative criterium. Moreover, the theoretical investigation of the vocabulary growth temporal patterns can be used as a technique for spam detection. As shown in Capocci et al. (2009b), the peculiar, sublinear relationship between age and number of tags of a resource is characteristic of a folksonomy, and deviation from this behavior are quite often a signal of spam activity.

SONY-CSL The development of NoiseTube is part of an entirely new class of applications (Patel-Predd, 2009; The Economist, 2009). The inclusion of folksonomies greatly improved the way in which the participants provided a bottom-up way for representing the issues of their exposure to pollution in a much more accurate way and at a semantic level, improving the navigation/search of a large amount of environmental data. Despite its young development, NoiseTube has already been presented at several workshops, conferences and covered by the press (c.f to final PDK document for a full list of these publications). Furthermore several collaborations with NGO (e.g. Awaaz foundation, India) and environmental agencies (e.g. BruitParif, the official observatory of Paris) are currently setting up to continue the research about this participatory approach after the TAGora project.

UNI KO-LD MyTag has been under active development until the end of the project and will also be used in the future to implement new features as a result of the ongoing research in this area.

UNIK BibSonomy will continue to exist after the end of TAGora, and aims to attract further users. The implemented results of the project will thus remain visible for a broad scientific audience. Follow-up projects that will ensure the existence of BibSonomy have already been started.

UNI-SOTON The Live Social Semantics application will continue to be developed in collaboration with the SocioPatterns.org project, and will be deployed at upcoming conferences, including ESWC2010. The methods we developed for tag filtering, tag sense matching, and inference of user interests will be carried forward in a newly funded EU project WeGov.

Chapter 2

Workpackage progress of the period

Following are the objectives that were planned for all the workpackages for the third year of activity of the project. For the tasks started in the second year see the file PA2.

2.1 Workpackage 1 (WP1) - Emergent Metadata

2.1.1 Objectives

Following are the objectives of the research carried out during the third year of the project.

The objectives of this WP were to collect the data necessary for carrying out all the research and investigations detailed in the other WPs, and to coordinate the efforts of collecting and storing the data whenever required. Much time and effort was spent by the consortium on data collection in the first year of the project which lead to gathering an impressive amount of very valuable data which was key for our research and analysis. Data collection continued over the second year of the project, but was more specifically targeted towards completing our existing data collections (e.g. Delicious and Flickr data) with additional information that was not collected in the first year, as well as collecting new data sets to further support our various research and development tasks.

As well as collecting data, we also encourage data sharing through the TAGora site where a page is available for describing our data collections and for pointing the public to where some of the data sets can be obtained from. The most valuable of those data sets are now available in RDF from the project website.

Data collection have naturally slowed down in year 3 of the project, and moved from mass collection to very specific crawling-on-demand, as demonstrated by the services from UNI-SOTON (section 1.2 and D4.5). Additionally, much data was collected during the Live Social Semantics experiments at ESWC09 and HT09 (see D4.5) which was used to make recommendations on conference talks and social contacts.

An update on data collection activities with respect to the four tasks of WP1 will be given below.

2.1.2 Progress

In this section the task responsibles must describe the progresses achieved during the third year of the project for each of the tasks described above.

Task 1.1 Data from collaborative tagging (folksonomies)

In order to test the Semantic Walker Model (see sect. 2.4.2 below) we considered the South

Florida Free Association Norms database (http://w3.usf.edu/FreeAssociation/) that includes roughly 60000 associations. Though this is not precisely a folksonomy database its adoption has been crucial as a proxy for a latent shared semantic graph. The South Florida Free Association Norms database has been compiled since 1973 by Douglas L. Nelson and Cathy L. McEvoy from University of South Florida and Thomas A. Schreiber from University of Kansas. The technique to collect the data strongly exploited human contributions. More than 6000 participants produced nearly three-quarters of a million responses to 5019 stimulus words. Participants were asked to write the first word that came to mind that was meaningfully related or strongly associated to the presented word on the blank shown next to each item. For example, if given BOOK, they might write READ on the blank next to it. This procedure is called a discrete association task because each participant is asked to produce only a single associate to each word. Here are the figures of the database: 72176 responses or targets given to a stimulus cue; 61880 responses marked as YES in the database. YES means that the target was normed, i.e., reused as cue; 10469 is the size of the target set; 5018 is size of the cue set; 4870 different words used both as cues and targets; 4845 is the size of the Strongly connected component (SCC) (i.e., the set of all the words among which all the possible pairs are connected by at least a possible path); 25 words have been used both as cues and targets but they don't belong to the SCC, instead to the IN component; 5599 words are in the OUT component: they have not been used as cues. (their number is the difference between the size of the target set and the size of the cuetargetset). The dataset is freely available trough the http://w3.usf.edu/FreeAssociation/ website.

Task 1.2 Data collection from the bibliographic reference sharing system BibSonomy

We continued publishing BibSonomy dumps every half year. For the ECML PKDD Discovery Challenges 2008 and 2009, we generated additional test and training data sets. The BibSonomy data were also used for the Viszards Session of the Sunbelt 2009 conference.

Task 1.3 Data from experimental tag-based navigation systems at Sony CSL

One of the datasets collected by the SONY CSL is the Armin Linke dataset. The dataset originated from an art installation called "Phenotypes/Limited Forms" that uses photos by artist Armin Linke and that has been on display at several exhibitions in Europe and beyond. The installation displayed 1.000 heigh quality photos, and visitors had the opportunity to create their own photo album out of these and tag it with a title. The dataset consists of around 190.000 Tag assignments made by around 23.000 users providing around 19.000 tags for 2.400 different Images. The number of images is higher then 1.000 since the set of images changed between different locations were the exposition took place.

Data collection for the NoiseTube application, the successor of Zexe.net, has also be initiated. Currently the data is directly available from the website (http://noisetube.net) via a web API either in CSV, JSON, KML or GeoRSS format. The data gathering started in April 2009. Because this is a relatively young project, the dataset currently contains only 10-20 users who provided about 20.000 noise measurements along with 400 tags.

Task 1.4 Collecting data from online recommendation systems UNI-KO-LD and UNI-SOTON

One of the services built by UNI-SOTON and used in MyTag as well as in LSS (see D2.5 and D4.5) is a Tag Sense Matching service, which recommends DBPedia.org resources and W3C Wordnet synsets that may correspond to the meaning of a given tag. To build this services, we collected processed all DBPedia pages to generate a large knowledge base, containing information such as the top 50 representative terms per page and their frequency of occurrence, list of unique words in each page, list of pages with specific words, etc. This data was rendered in RDF and stored in the

4Store[1] triplestore. In summary, TFIDF of the whole of Wikipedia was generated and used for tag sense recommendations.

Task 1.5 Public delivery of documented data collected by the Consortium

This task is concerned with ensuring the publication of the main TAGora datasets in reusable formats. To this end, we have now released several of our datasets in RDF. These datasets can now be freely downloaded from the TAGora website. Deliverable 1.4 briefly describes those datasets. In some case we decided to anonymize the names of the account holders to preserve their privacy whenever possible.

2.1.3 Deviations and Corrective Actions

PHYS-SAPIENZA: none
SONY-CSL: none
UNI KO-LD: none
UNIK: none
UNI-SOTON: none

2.1.4 Deliverables and Milestones

Del. No.	Deliverable name	WP No.	Date due	Actual/ Forecast delivery date	Estimated indicative person-months	Used indicative person-months	Lead contractor
1.4	(T1.5) Public delivery of data collected by the Consortium and related documentation (Month 38).	1	Oct. 15th 2009	Sept. 15th 2009	4	4	**All**

2.2 Workpackage 2 (WP2) - Applications

2.2.1 Objectives:

The objective of WP2 for the last year of the project was to close the cycle from the applications to models and back again. This meant especially that the best results of WPs 3 and 4 should be implemented in the systems. A second objective was to lead the software engineering process of the state such that the systems can be used further after the end of the project. Depending on the system, this may mean different things, ranging from cleaning-up the source code and putting it under an open source licence to improving the features of the online version to attract more users.

[1] http://4store.org/

Task 2.1 Social tagging for online scientific communities:

Task 2.1.1 - Folksonomy web site for sharing of bibliographic data

BibSonomy serves as platform for our experiments, and helps thus for closing the loop from the expermiments back to the systems. Beside the implementation of some additional features to attract more users and keep the existing users in the system, our focus of the last year were frameworks for spam detection, tag recommendations, user recommendations, and logging.

Task 2.1.2 - Folksonomy peer-to-peer system for sharing of bibliographic data

The development of the peer-to-peer tagging system Tagster was facing technical problems and the aquiration of a critical mass of users was difficult. This resulted in a significant delay with respect to the original schedule. Hence, following the comments of the reviewers to divert the efforts, originally dedicated to Tagster, to other important tasks we focused on the extension and improvement of MyTag (http://mytag.uni-koblenz.de/). This included an improvement of the result ranking, better usability and other optimizations of the MyTag implementation.

Task 2.2 Tag-based navigation systems (music and image data).

The Ikoru system evolved out of a research project to explore how the tag-based navigation could be improved with the use of data analysis. We also wanted to be able to register the detailed browsing history of the visitors in order to analyse possible relations between navigation, tags, and contents. We aimed to develop a reusable, open software platform using Web standards so that it can be deployed in specific studies or extended with new features.

2.2.2 Progress

In this section the task responsibles must describe the progresses achieved during the third year of the project for each of the tasks described above.

Task 2.1 Social tagging for online scientific communities:

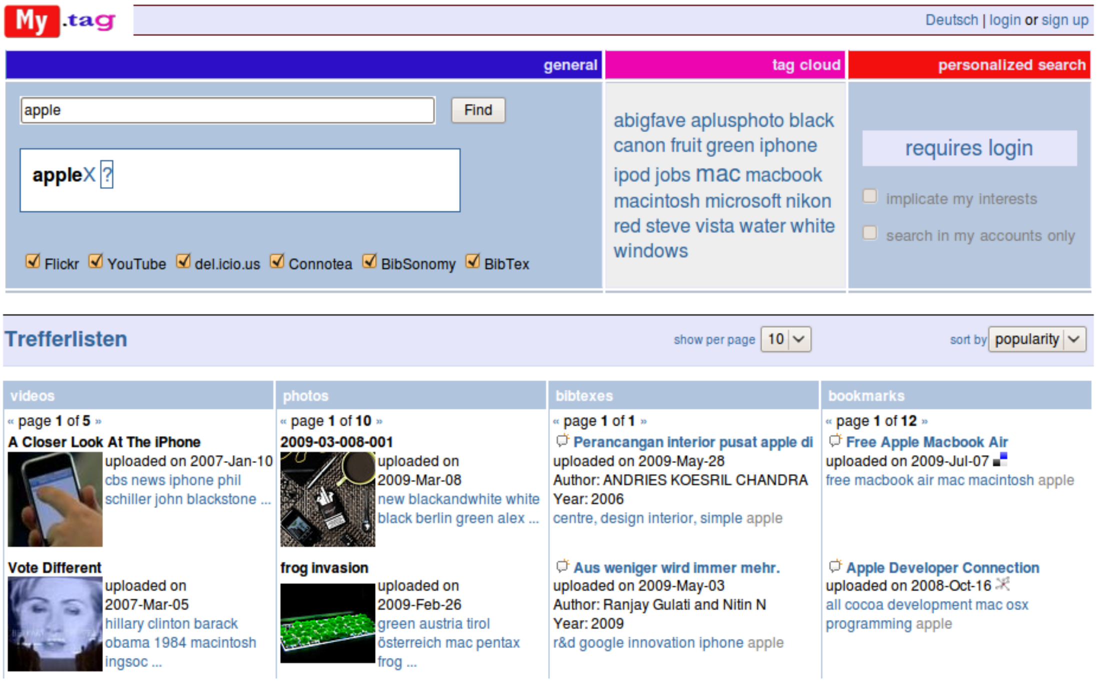
Task 2.1.1 - Folksonomy web site for sharing of bibliographic data

The functionality of BibSonomy was further enhanced, as described in Deliverable 2.5. In particular, we implemented a logging framework and a spam framework, and set up the design of the recommender framework. The latter had to be implemented outside of TAGora, financed by a national follow-up project, due to the limited TAGora resources. The frameworks were the technical platform for the experiments performed in WP 4.

Task 2.1.2 Folksonomy peer-to-peer system for sharing of bibliographic data

The effort formerly put into the development of Tagster has been diverted to other important task like the improvement and extension of MyTag (http://mytag.uni-koblenz.de/). MyTag is a cross folksonomy search and recommendation tool which has been developed within the context of TAGora, and is being used to disseminate the project through the software implementation of theoretical research. It allows for searching different content types like photos, videos and bookmarks in parallel. For this purpose, it queries the public APIs of different tagging systems and presents the retrieved results in a separate column for each content type (see Fig. 2.1). Besides the search across different tagging systems, it collects the search interests of registered users and offers them a personalized ranking of results. Furthermore, it now also offers an intelligent search assistant that helps in disambiguating the current search terms by grounding them to possibly relevant articles found in DBPedia.

Altogether, the following improvements were introduced to MyTag during the last year:


My.tag
Deutsch | login or sign up
general
tag cloud
personalized search
apple
Find
appleX ?
☑ Flickr ☑ YouTube ☑ del.icio.us ☑ Connotea ☑ BibSonomy ☑ BibTex
abigfave aplusphoto black canon fruit green iphone ipod jobs mac macbook macintosh microsoft nikon red steve vista water white windows
requires login
□ implicate my interests
□ search in my accounts only
Trefferlisten
show per page 10
sort by popularity

videos	photos	bibtexes	bookmarks
« page 1 of 5 »	« page 1 of 10 »	« page 1 of 1 »	« page 1 of 12 »
A Closer Look At The iPhone uploaded on 2007-Jan-10 cbs news iphone phil schiller john blackstone ...	2009-03-008-001 uploaded on 2009-Mar-08 new blackandwhite white black berlin green alex ...	Perancangan interior pusat apple di uploaded on 2009-May-28 Author: ANDRIES KOESRIL CHANDRA Year: 2006 centre, design interior, simple apple	Free Apple Macbook Air uploaded on 2009-Jul-07 free macbook air mac macintosh apple
Vote Different uploaded on 2007-Mar-05 hillary clinton barack obama 1984 macintosh ingsoc ...	frog invasion uploaded on 2009-Feb-26 green austria tirol österreich mac pentax frog ...	Aus weniger wird immer mehr. uploaded on 2009-May-03 Author: Ranjay Gulati and Nitin N Year: 2009 r&d google innovation iphone apple	Apple Developer Connection uploaded on 2008-Oct-16 all cocoa development mac osx programming apple

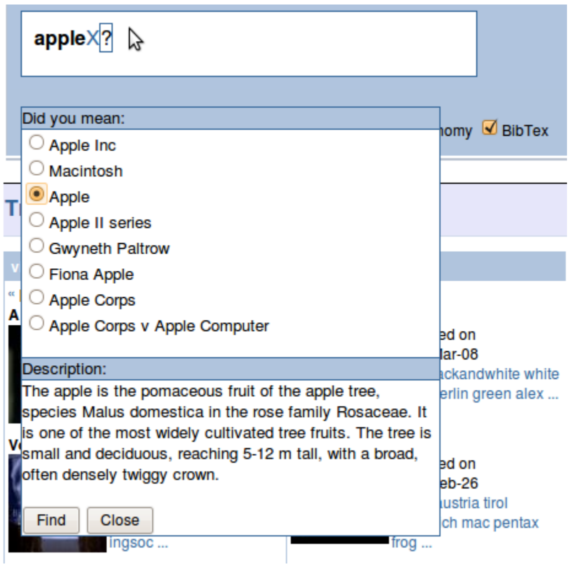
Figure 2.1: User interface of MyTag. The upper part contains the search interface, showing the current search terms and the tag cloud of all terms associated with the current search results. The lower part contains the actual result lists. Each content type is shown in a separate column.

- Incorporating further platforms and content types: With BibSonomy and Connotea, two additional platforms were introduced into the system. By incorporating BibSonomy, MyTag now also supports the search in bibliographic references.

- Merging search result lists for the same content type: By incorporating BibSonomy and Connotea, we now retrieve search results for bookmarks from three different platforms. This required to introduce an algorithm into Mytag that merges the bookmarks coming from Delicious, Connotea and BibSonomy into a single result list. The technical details about the merging algorithm are available in Grabs (2009).

- Intelligent search assistant I: In a collaborative effort with the Southampton team, we introduced an intelligent search assistant into MyTag. It automatically analyzes the current search terms of the users and sends it to a disambiguation web service offered by the Southampton University. This web service grounds the search tags in articles from DBPedia and returns possible related terms. MyTag then analyzes the returned list of possible meanings of the search terms and filters those meanings which are not represented in the search results retrieved from the tagging platforms. The remaining grounded terms are then presented to the user (see Fig. 2.2). The user can then select the intended meaning of the search term and re-rank the current list of results so that resources corresponding to the intended meaning are ranked higher. A description of the approach is published in Dellschaft et al. (2009).

- Intelligent search assistant II: In Abbasi and Staab (2009) a method is proposed how to identify generalized and specialized tags. This analysis was applied on the collected Flickr data set and the resulting lists of generalized and specialized tags were also used for suggesting tags to a user during his search. But the evaluation in Scharek (2009) showed that users only seldom found the suggested tags useful for refining their search. Because of these mixed evaluation results, this search assistant was not included into the publicly available version of MyTag.



appleX?

Did you mean:

○ Apple Inc
○ Macintosh
● Apple
○ Apple II series
○ Gwyneth Paltrow
○ Fiona Apple
○ Apple Corps
○ Apple Corps v Apple Computer

Description:

The apple is the pomaceous fruit of the apple tree, species Malus domestica in the rose family Rosaceae. It is one of the most widely cultivated tree fruits. The tree is small and deciduous, reaching 5-12 m tall, with a broad, often densely twiggy crown.

Figure 2.2: Interface of the search assistant for disambiguating the current search terms. The search assistant pops up if the users clicks on the question mark next to the search terms. For each of the found meanings, a short description based on the DBPedia article is presented to the user.

MyTag Data Set The MyTag data set contains information about all queries submitted to MyTag and the results on which the user clicked. The search queries are grouped into search sessions, i. e. subsequent queries done by the same user can be associated with each other. The MyTag data set is completely anonymized, i. e. it contains no information about who did a single search query regardless whether it was done by a registered user or not.

The size of the MyTag data set is available in Tab. 2.1. The following information is available in the data set:

- tags (id, label): A list of all known search terms with their internal ID and the natural language label.
- search-activities (id, date, sort-by, search-in, session-id): Each row corresponds to a single search request. If a user made several requests without closing the browser, all of them have an identical session-id.
- search-activities-tags (search-id, tag-id): This table contains the search terms that were used for the different search requests.
- tagging-activities (id, date, system-id, resource-id, search-id): Each row corresponds to a single search result on which a user clicked. It can be associated to a concrete query by its search-id. The system-id gives the tagging system from which the result was retrieved and resource-id is the system specific identifier of the resource (e. g. a Flickr ID in case of photos or a URL in case of bookmarks).
- tagging-activities-tags (tagging-id, tag-id): This table contains the tags that were assigned to the search result on which the user clicked.

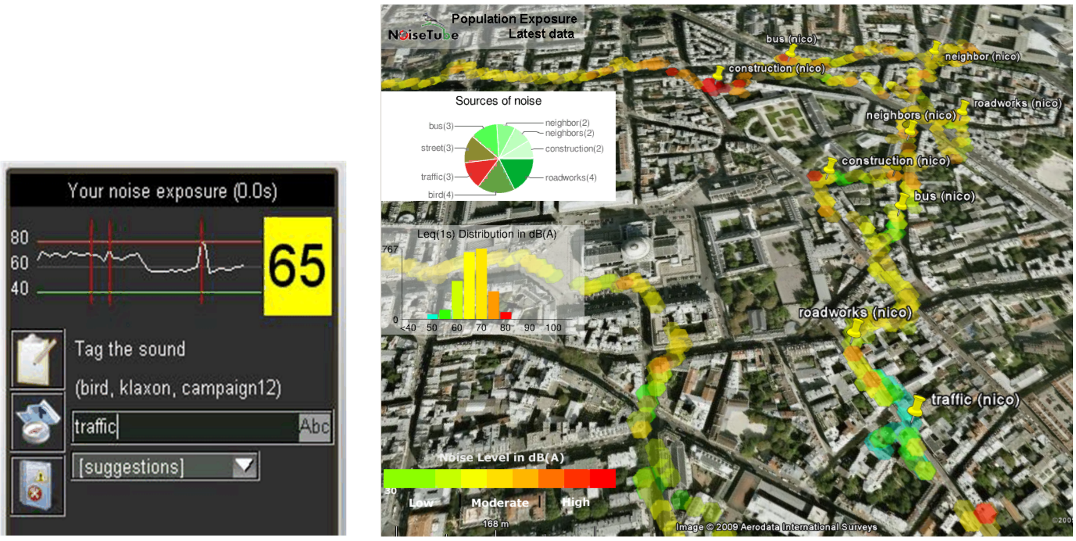
Task 2.2 Tag-based navigation systems (music and image data).

For Ikoru,in the second Periodic Activity Report SONY CSL declared its decision not to promote the website to a large audience. The reason for this decision was that maintaining an active multimedia Web site would require too many resources. Taking into account the reviewers comments,

tags	search-activities	search-activities-tags	tagging-activities	taggging-activities-tags
5461	2935	3777	715	9147

Table 2.1: Size of the MyTag data set (August 2009).

SONY CSL decided to provide a lightweight API for Ikoru to make the main feature such as the hybrid classifier mixing tagging and low-level features for images accessible to Flickr. Turning Ikoru into a web service with original features for Flickr increases its chance to attract a larger audience. The Zexe.net project focussed on exploring new applications of tagging for the benefit of off-line communities facing a variety of issues related to the sustainable exploitation of a commons. Through the use of smart phones, communities in different cities around the world have published content on the Zexe.net platform, resulting in so-called Community Memories Steels and Tisselli (2008) which help these communities represent and raise awareness about shared concerns. Starting from the 3rd year of the TAGora project we began to extend this idea by applying collaborative tagging to a new type of resource: the exposure of individual people to pollution. The Zexe.net platform was consolidated and extended to the case of noise pollution and was renamed to *NoiseTube*. NoiseTube is a participative sensing Burke et al. (2006) and tagging platform that aims to enable citizens to measure gather, manage, visualise their exposure to noise pollution in their everyday life. (see the delivrable 2.5 for more information)



(a) The user interface of the mobile application allowing to measure, tag and send the current exposure to noise (in dB(A) to the web application to inform the community and build maps

(b) The visualisation of a noise exposure map in a district with a semantic layer to identify the sources of noise thanks to tagging features

Figure 2.3: Screenshot of NoiseTube

2.2.3 Deviations and Corrective Actions

In case there is any deviation from the project objectives it must be described here by the task responsibles:

SONY-CSL: SONY-CSL decided to extend the Zexe.net principles to the NoiseTube project. Like Zexe.net, the focus remains on supporting offline communities with new applications of collaborative tagging in real world situations, in particular for sustainability-related issues, by using mobile phones. This deviation from the original objective of tagging system for multimedia resources to the exposure of individual citizens to noise pollution. allowed SONY-CSL to find new tagging usage and new ways to disseminate research on tagging to real-world and sustainibility-related fields.

UNI KO-LD: The development of Tagster was facing intricate technical problems during the second year and was thus lagging behind the original schedule. The reviewers also argued that reaching a critical mass of users under these circumstances until the end of the project will be difficult. So they advised to divert the effort on other task which which were at that point underachieving. Thus, the work on Tagster was stopped and more effort has been put in the developments of additional functionality for MyTag, like disambiguation and improved ranking.

UNIK: none

2.2.4 Deliverables

Del. No.	Deliverable name	WP No.	Date due	Actual/ Forecast delivery date	Estimated indicative person-months	Used indicative person-months	Lead contractor
D2.4	(Task 2.2) Final version of the Tag-based navigation system for images (Month 38).	2	Oct. 15th 2009	Sept. 15th 2009	2	2	**SONY-CSL**
	(Task 2.2) Final version of the Tag-based navigation system for music (Month 38).	2	Oct. 15th 2009	Sept. 15th 2009	2	2	**SONY-CSL**
D2.5	(Task 2.1) Final report on tagging systems update and usage (Month 38).	2	Oct. 15th 2009	Sep 15th t. 2009	2	2	**UNIK (ALL)**

2.3 Workpackage 3 (WP3) - Data analysis of emergent properties

2.3.1 Objectives

Following are the objectives of the research carried out during the third year of the project.

The goal of the third year of the project was to further improve the analysis methods, e.g. community identification, network analysis, emergent metadata statistics, and cross-folksonomy analysis. As mentioned by the reviewers, the analysis should also include results at the micro level, not only at the macro level. Furthermore, the feedback into WP2 should be intensified by integrating more methods in the developed applications. This includes methods for automatic tag classification and search improvements, in terms of precision and recall, by considering semantic similarities and correlations of tags and resources. If a method is not directly implementable in our applications there might be the option to expose its functionality as a web service, e.g. the processing steps of cross-folksonomy analyis like filtering an mapping.

Task 3.1 Emergent metadata statistics

Task 3.1 deals with the quantitative statistics coming from the analysis of the datasets provided by WP1. While in the first year we mainly analyzed static statistical properties, during this and previous year we focused our investigation on dynamical statistical properties in order to better characterize the cooperative behavior of users, if any.

In particular, during this year we focused our attention on the bursty tagging activity occurring in a folksonomy in order to understand whether different bursty events, which are the result of the apparently uncorrelated action of users, might lead to any loose tag categorization. We also tried to understand, by using the tool of the Inverse Partecipation Ratio, whether users reach a spontaneous consensus on the description of resources in terms of significant tags.

Task 3.2 Network/graph analysis:

The objective of this task is to provide the necessary methods for analyzing and describing the topological and dynamical properties of the complex networks occurring in tagging systems. The analysis of these networks is essential for the theoretical description of the system but they may also be used for more practical purposes like the inference of tag semantic classes, which could improve the ease of navigability of folksonomies.

We studied whether social interaction carries any influence on the semantic relatedness between users, i.e. whether users supposed to be in close social relationship are also prone to use more similar tags. To this aim, we analyzed the evolution in time of the semantic assortativity in the social networks hosted by the `Flickr` folksonomy, based both on the contact data and on the group membership data provided by the users themselves.

Moreover, in order to understand whether tag order in a post carries any semantic meaning we also analyzed the directed version of the cooccurrence network of tags defined last year.

Task 3.3 Cluster/community identification

This task concluded with Deliverable 3.2 at the end of the second project year.

Task 3.4 Semantic Inference

When annotating resources in folksonomies users tend to assign only a limited set of tags of their choice. They might not add many relevant tags to the resources. This results into sparseness of data and makes it difficult to search relevant resources, especially when there are only few resources in the folksonomy relevant for a combination of query tags. This problem can be overcome by enriching the vector space model, which is commonly used for the retrieval and ranking of results, with the missing tags. One way to infer the missing tag relations is the combination of resource, tag, and user co-occurrences.

Task 3.5 Cross-Folksonomy Networks

This task aims at bridging between multiple folksonomies to facilitate researching tag evolution across folksonomies and cross-folksonomy recommendations. Reaching these goals requires an understanding of how tags from different folksonomies or from individual tag clouds can be filtered, disambiguated, and mapped to each other. To this end, much effort was spent over the last two years of the project on researching methods for achieving this data integration, and on developing the infrastructure and services for crawling and mapping data from multiple folksonomies.

Task 3.6 Collaborative tagging and emergent semantics:

Task 3.6a Improving Navigation for Images

As part of the process to improve the navigation within the Ikoru application for images a closer analysis of related datasets from the deployed applications was initiated. Specifically the analysis of the data collected through the "Phenotypes/Limited Forms" installation. In this installation museum visitors were engaged in tagging in a physical space, and the exhibition was a mere extension of the virtual Ikoru Interface into the real world. In this configuration no feedback back exists from the Ikoru system to the artist or the visitors of the exposition. The objective of the data analysis is therefore to close this gap in the communication loop and provide feedback from the virtual tagging-application of Ikoru to the real world tagging application.

Task 3.6b Improving Automatic Classification of Music

The objective of this task is to study the relation between tags and content-based analysis. Is it possible to ground tags? Is it possible to improve the navigation based on tags with data extracted from the content? Can we reduce some of the limitations of tagging, such as the problems of homonomy and synonymy? And vice versa, can tags offer a support for automatic classification schemes These are some of the questions that we aim to address.

2.3.2 Progress

In this section we describe the progresses achieved during the third year of the project for the tasks described above.

Task 3.1 Emergent metadata statistics

The emergence of topics in folksonomy evolution has been investigated by studying the statistics of bursts in the tagging activity of several folksonomies. A reasonable hypothesis would predict that topics, mimicked by tags, maybe divided in two class according to their usage: on one hand, bursty tags related to special events or other peak of interest (e.g. “worldseries2008”), and regularly used one, related to daily activities of users (e.g. “blog”). If this was the case, the categorization of topics could be performed (also) based on the dynamics of the corresponding tags. Inspired by methods employed in the theoretical modelling of the physics of earthquakes, we have explored in detail the statistical distribution of interarrival times of tag occurrences.

Such distribution reflects time correlation in tag usage. Bursts of activity should map into a large number of short interarrival times, and a few long ones. By contrast, if tags were assigned randomly, with a fixed probability at each time step as in a Poissonian process, their interarrival time distribution should follow an exponential function. Similar analysis have already been performed for other data sets, namely texts, showing that the distribution of word occurrence is not random and deviation from a Poissonian picture are present. Observed fat tails in the distribution of word interarrival times have been detected in texts, and put into relation with the underlying semantics. Tag stream, too, display such power–law behaviour in the interarrival time distribution. Such properties is displayed by both frequent and rare tags.

A similar scientific problem has been faced by geologists studying earthquake events: apparently, the bursty dynamics of extreme events is clearly distinct from dynamics of small events, the latter being almost a continuous one with hundreds of small earthquakes observed everyday. However, a detailed statistical analysis revealed that large and small earthquake follow the same dynamics, when time scaled are opportunely tuned. In brief, the dynamics of large earthquake is similar to the one of small ones observed at a shorter time scale. Hence the same uncertainties affect small and extreme earthquake events Bak et al. (2002).

By a similar scaling analysis, it is possible to observe that the tagging activity related to popular topics follows the same dynamics of the activity related to more specialized ones, if a suitable rescaling of the time unit is performed. As a consequence, by a mere observation of the tagging dynamics it is not possible to clearly discriminate the character of tags (bursty as opposed to

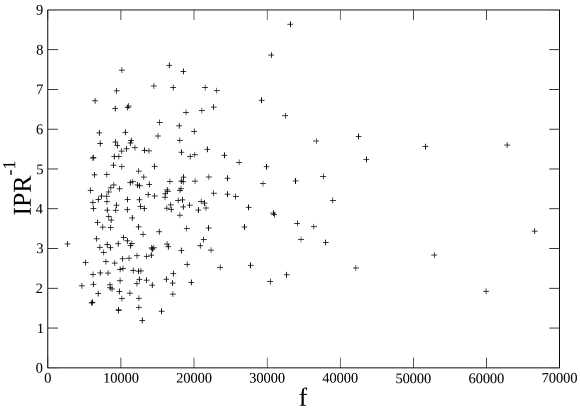


Figure 2.4: The IPR of tag streams associated to individual internet pages, as a function of their number of tag assignments f.

regularly used ones) and, hence, infer their categorization.

The presence of complex time correlations in tags cannot be explained by semantics in itself, as happen in written texts. Nevertheless, as pointed out by statistical physics, long-tailed time correlations are often related to cooperative phenomena and collective dynamics taking place in the system. In a folksonomy, hence, such statistical properties could be naturally linked to social dynamics.

This hypothesis has been tested by analysing whether users in the largest folksonomy examined, Del.icio.us, tend to reach an agreement on the description by tags of resources. On one hand, it had already been shown that the number of distinct tags assigned to a resource grows in time at a sublinear pace Cattuto et al. (2007). This would suggest that annotations by new users continuously add new features to the description of the considered resource. Such proliferation of tags referring to a single resource, however, is not meaningful in itself, since it includes typos, translations, synonymes etc.

So, we have measured the number of significant tags only assigned to each resource. For each resource, we have measured the Inverse Participation Ratio (IPR) of its tag cloud. The IPR of a resource is equal to $IPR = \sum_s (\frac{f_s}{\sqrt{\sum_t f_t^2}})^4$, where f_t is the number of assignments of the tag t to the resource. By this method, we have shown that the number of significant tags is very small even for resources annotated by thousands of different users. The IPR of resource tag clouds remains in the range of values between 1 and 10 for all resources included in our database, as shown in figure 2.4. This shows that, although the absolute number of tags assigned to a resource grows with time, users nevertheless focus their description on a very limited number of tags, which remains in a characteristic range for all resources. Results are detailed in ref. Capocci et al. (2009b).

Task 3.2 Network/graph analysis:

Task 3.2.1 - Topological properties

Folksonomies embed several explicit and implicit networks. Examples of implicit networks are the co-occurrence networks where, for example, links are drawn between tags used in the same post. Some folksonomies, such as Flickr, also embed explicit social networks, since users can express which friends they are in contact with and which groups of interests they are part of.

We have examined whether social interaction influences semantic relatedness between users. To understand this, we have defined several graph based on the social relationships established on Flickr. We have studied graphs where links are a signature of mutual or directed friendship between users.

The friendship network is a scale-free one, with a power law distribution of the degree among users. A similar feature is observed in the distribution of the number of groups a user belongs to. Another typical feature of social networks has been observed in the Flickr social network, that is, assortativity. In fact, it has been shown that high–degree users are preferably friends of other high–degree users. Analogously, users tend to be friends of users with similar number of group memberships and tagging activity (both in number of used tags and of tag assignments).

Then, the social network has been compared with the semantic similarity between users. Two kinds of semantic relatedness have been adopted. First, the number of shared tags in the tag clouds of users. Second, the cosine similarity between the vectors representing the tag clouds, where each component is equal to the frequency of a tag in the user vocabulary, and the similarity is the cosine of the angle between the two vectors. Such similarity measures have been compared with the distance on the Flickr social network, defined as the length of the shortest path between the two nodes.

Both similarity measures confirm that friendships and semantic similarities are strongly correlated. Similarity decreases for distant nodes according to both measures. However, such results have to be correctly interpreted by the comparison with a suitable null model where tags are re-assigned by maintaining the global tag frequency distribution and the number of tags used by each users, but destroying local correlation. The number of shared tags, for example, is larger in neighbor users but is no signature of semantic similarity. In fact, in the null model, too, one observes the same relation between the network distance and the semantic relatedness.

By contrast, the cosine similarity displays a clearer signature of the influence of social interaction on semantic similarity. While in the real dataset the semantic similarity decreases for larger distance between the users in the social network, in the null model there is a very weak dependence between the two quantities. Details on this analysis can be found in ref. Schifanella et al. (2009).

Users annotate resources with posts by adding tags in a certain order. The tag-tag cooccurrence graphs studied so far, do not consider tags order inside posts. In order to establish whether tag ordering has a kind of relevance, we construct and analyze a weighted, *directed* cooccurrence network that fully encodes the order of tags inside posts. A directed link is drawn from one tag to its following tag.

In order to understand whether tag ordering in posts is a semantic relevant feature and not the result of a random process, we exchanged randomly the position of tags inside single posts and create a new fictitious folksonomy. This simple randomization process will result in a directed weighted tag-tag cooccurrence graph where correlations among tags in the same posts are artificially destroyed.

We then measure the correlation between in-strength and out-strength of nodes with a scatter plot and find that the shuffling process narrows the original picture, thus showing that the order of tags in posts may be important and not the result of a random process. We conjecture that the average user annotates resources by using tags with increasing (or decreasing, it is not clear at this stage) degree of generality in the posts. Details can be found in Benz et al. (2008).

Task 3.2.2 Dynamical properties

3.2.2a social and semantic relatedness

The relation between social and semantic relatedness has been examined during its time evolution too. A set of measurements similar to the one reported in the previous section have been performed on subsequent snapshot of the Flickr folksonomy, to monitor whether the results are robust and stable in time. Beside the friendship network, we have also focused our attention on the social network based on group memberships. In such a class of graphs, users are connected if they belong to the same groups, or to a sufficient number of common groups.

To study the evolution of semantic similarity and social interaction, we have divided a whole year of the Flickr dataset into slice of fixed time length, and limited our analysis to the users active in all time interval. This subset of users has been followed during the evolution of Flickr, by studying the tag clouds relative to each time interval, and their similarities by several methods. We have measured the average semantic similarity of neighbor users as a function of time, of the chosen network and of the adopted measurement method. As done in the static analysis mentioned above Schifanella et al. (2009), we have compared the data with a suitable null model, where all statistical properties of the tag stream are maintained, but the association between tags and users have been randomly reshuffled.

As in the previous study, the social linkage between users is found to correspond to a higher semantic similarity during the whole data set explored. All measures adopted (cosine similarity, TF-IDF methods, shared tags) to observe the alignment of tag clouds show that the similarity between the social network and the semantic network does not develop gradually, but rather is quite a steady property of the Flickr social network since its beginning.

A hierarchy of social interaction intensities has therefore been established. By comparing the average semantic similarities in all social networks, one observes that the friendship social network corresponds to a much stronger semantical relatedness, even greater if one limits the analysis to the social network built on mutual friendships. A shared group membership, on the other hand, corresponds to a weaker linguistic interaction. Even for users sharing 10 groups or more, the semantic similarity remains well below the values found in the friendship–based networks Capocci et al. (2009a).

3.2.2b group dynamics

Some analysis of social network dynamics in folksonomy has been carried on, in particular on flickr, where several social networks are explicitly exposed, as mentioned above. In particular:

1. `Flickr` contacts: `Flickr` each `Flickr` user maintain a list of “friends”, or contacts, which is public; this defines a first explicit directed social network;

2. `Flickr` groups: `Flickr` users can subscribe to Groups, which are comprehensive of a Pool of adherents’ photos, and a discussion forum; many of this groups are public, as well as the list of their members; this represent a bipartite network (user,groups), which can be eventually projected in a social network between users.

During the flickr crawl, data about groups and their memberships have been collected, and this allows for a first series of statistical analysis, which feeded the discussion of a working group at the Dagstuhl workshop [Baldassarri et al. (2008)]. In particular a first analysis of the statistics of the network revealed a typical scale free topology scenario. In Fig.2.5, we show the distribution of users per group (number of members), as well as the number of groups per user (number of affiliations). Moreover, given a couple of users, we considered the number of groups they have in commons, and the corresponding distribution is also displayed in Fig.2.5. However, reshuffling the memberships in an appropriate null model seems to show that there is a weak correlation between users affiliation. In order to better investigate this fact, we also try to cross the information from contact networks with group affiliations.

In particular, in collaboration with Diego Taraborelli (also present in Dagstuhl) and Camille Roth, we performed a study trying to relate the dynamics of `Flickr` groups with the social network of `Flickr` user contacts [Taraborelli et al. (2009)]. In Fig. 2.6 the concept of “group centered” social network is descriped: i.e. the social network of flickr contacts restricted to the members of a given group. The topology of this social network, as well as other information about the group, has been taken in consideration in order to preview the dynamics of the group, i.e. its growth rate. An example of the methods and results of the analysis, we show in Fig. 2.7, the growth rate of population (members) and content (number of pictures in the pool) versus a “reprocity index”, measuring the fraction of reciprocal friendship relations between members of the group. In

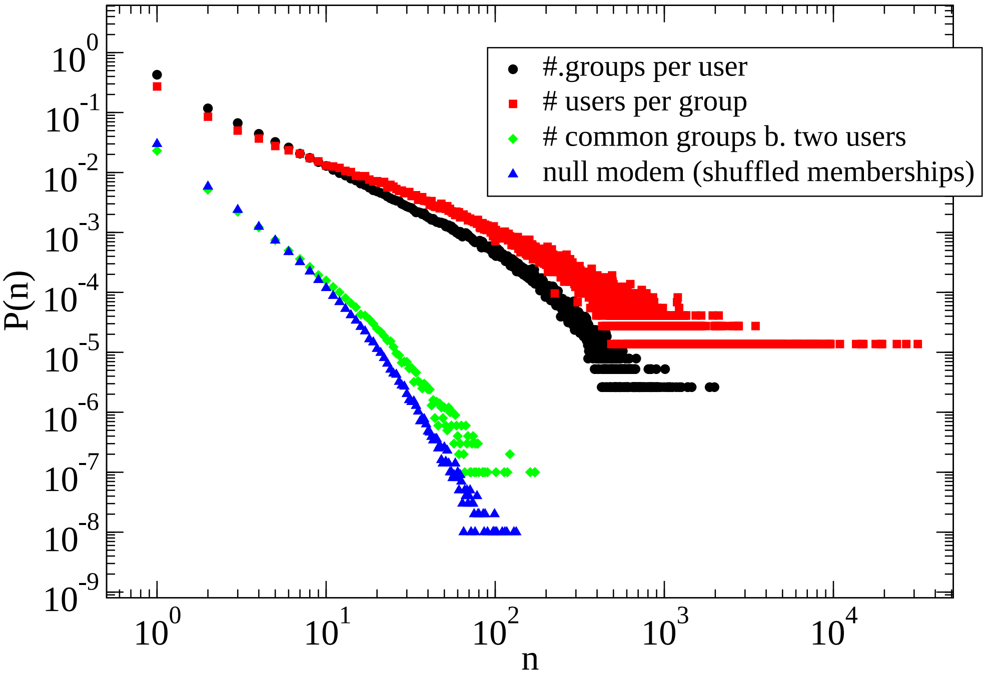


Figure 2.5: Group membership statistics.

general these kind of analysis show some weak, but not striking, correlations and further work is in progress.

Task 3.3 Cluster/community identification

This task concluded with Deliverable 3.2 at the end of the second project year.

Task 3.4 Semantic Inference

One Problem with folksonomy data is sparseness because users tend to assign just a few tags to resources. The *enRiched Vector Space Models for Folksonomies (RichVSM)* Abbasi and Staab (2009) tries to overcome this problem by inferring missing tags. For example a user can upload a funny image of 1970s and add the tags *funny* and *seventies* to it. Users can search this image by giving the tags attached to it. The users may tag resources with keywords of their choice. They might not add many relevant tags to the resources. This results into sparseness of data and makes it difficult to search relevant resources. Especially when there are only few resources in the folksonomy relevant for a combination of query tags. For example if a user is searching for funny pictures of 1970s using the tags *funny* and *seventies*. He will get only the images tagged with *funny* and *seventies* or ones that contain one of the two tags. The user will be unable to get resources that are tagged with *1970s* and *funny* but are not tagged with *seventies*, although the resources tagged with *1970s* and *funny* might be of interest to him. We hypothesis that there are many resources in folksonomies which are not searchable because they do not contain most of the relevant tags.

Using enriched vector space models, we show that one can find meaningful relationships between

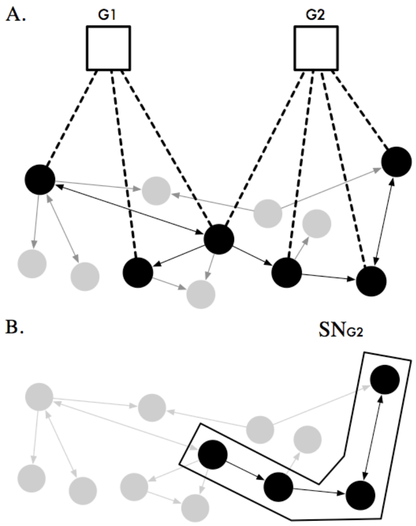


Figure 2.6: Definition of Group centered social network.

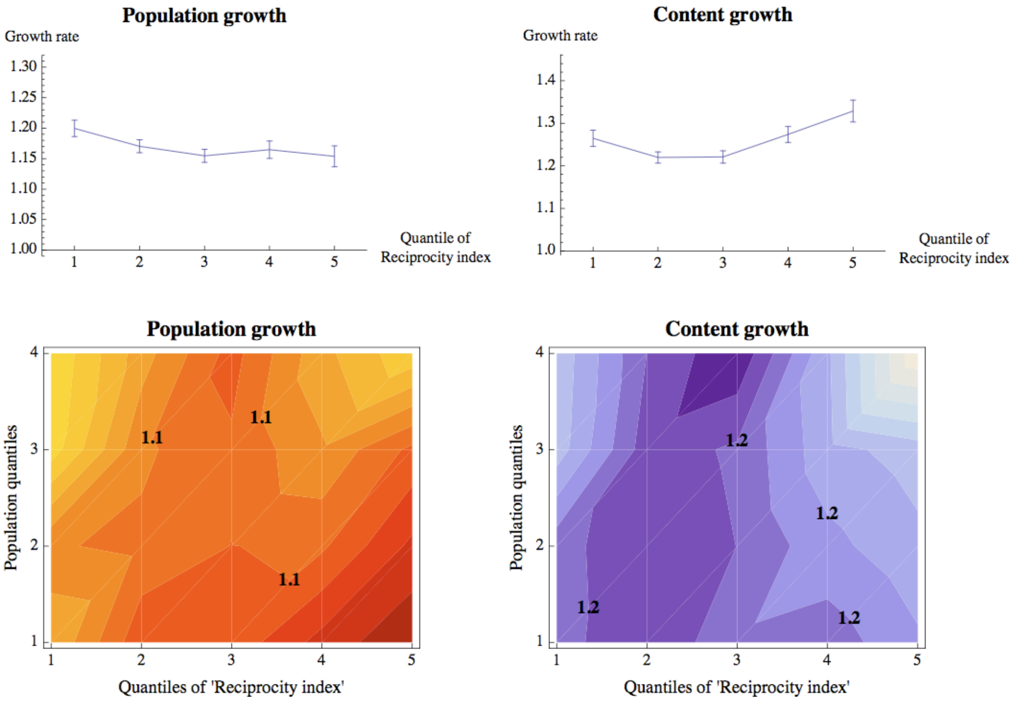


Figure 2.7: Group growth rates vs reciprocity index of group centered social network.

tags and then use these relationships to reduce the sparseness in folksonomies. We find the relationships between tags based on two dimensions, first the context of the tags and second the distribution of tags. We consider two types of tag contexts, the resource context (which resources are assigned a particular tag), and the social context (which users have used a particular tag). The resource context of tags helps in finding tags which are mostly used in similar kind of resources, whereas the social context finds broad relationships between tags based on the users' interests (represented by the tags they use). We also exploit two kinds of tag distributions, 1) similar tags and 2) generalized tags. We find relationships between similar tags by using the existing *cosine* similarity measure and propose a modified overlap coefficient to exploit generalization relationships between tags. We hypothesize that the statistic description of resources that use common tags exhibits different behavior than the statistic description of resources with uncommon tags. To test this hypothesis, we split the queried tags into three sets; having 1-10 search results, 11-50 and more than 50 search results respectively and perform experiments on these sets of queries. We also propose a method *Best of Breed* (BB), which selects appropriate enrichment model based on the number of relevant resources related to the queried tags. Experimental results based on a large scale evaluation (150 queries evaluated on a dataset of ~27 Million resources by 18 expert users) show that the enrichment of existing data by exploiting semantic relationships among tags helps in improving the search results, particularly for the queries which have a few relevant resources in the original data.

User Study Our large-scale dataset[2] was obtained by systematically crawling the Flickr system during 2006 and 2007. We filtered our dataset by removing those tags which were used by less than 10 users. Those users and resources were also removed from the dataset which did not use any tag. In the final dataset, we had data of about 27M photos, 0.3M users, and 92K tags. The exact statistics of the dataset are shown in table 2.2. We did all our experiments on this dataset.

users	tags	resources	tag assignm.
317,260	92,460	26,801,921	94,499,112

Table 2.2: Flickr filtered dataset statistics

For evaluation, we used the AOL query log (details in Pass et al. (2006)) which originally contained 20M queries from 650K users during three months from March to May 2006. Out of these 20M queries, we selected queries having 2 to 5 words for which the user had clicked on a link to the Flickr website. We split the queries into three sets, each set having 1 to 10, 11 to 50, and more than 50 exact matches (resources having all the queried tags) in the original vector space model. We randomly selected 50 queries from each of these three sets, resulting into 150 total queries for the evaluation.

We performed a user study for evaluating proposed and simple vector space models. The results were evaluated by 18 expert users (mostly PhD students) who were well familiar with search and image search. Each user was shown a search result page. The query was shown at the top of each evaluation page with images retrieved as a result. The title of the image was shown at the top of the image, tags on the right side, and evaluation options at the bottom of each image. Every user was given a set of queries and results obtained using different vector space models. Users were unaware of the method used for creating the search result page. Users were asked to mark an image as very relevant or relevant if the image matches the query, mark as don't know if they are not sure about the image, irrelevant or very irrelevant if the image does not match the given query. Queries were randomly distributed among users. The images marked as relevant or very relevant were considered as relevant and others as irrelevant in final evaluation. The results and

[2] The reference data set used for this evaluation is available at http://www.uni-koblenz.de/~goerlitz/datasets/tas_flickr.gz

analysis of the user study can be found in Abbasi and Staab (2009).

Task 3.5 Cross-Folksonomy Networks

As mentioned earlier, this task is concerned with building the tools for integrating data from multiple folksonomies. In year 2 of TAGora, we investigated how to filter tags, and how to ground them to URIs to build integrated semantic networks that cover multiple folksonomies data. Integrating this data creates a network of tags, users, and resources. This data was used for computing tag-cloud similarity of individuals across multiple folksonomies (Szomszor et al. (2008b)). We also investigated how such data can yield information about the interests of users, that could be scattered over several folksonomies (Szomszor et al. (2008a)).

In year 3 of the project, we concentrates our efforts on developing further, and in some cases opening access to, the services for cross-folksonomy integration and analysis. These services include:

- **Tag Filtering:** When people tag resource, be it a web page, photo, song, or video, they are free to choose any tag(s) they please. While it has been shown that this uncontrolled behaviour does result in meaningful semantic structures, the tag-clouds of particular individuals often contain misspellings, synonyms and morphologic variety. As a result, important correlations between resources and users are often lost simply because of the syntactic mismatches in the tags they have used. The Tag Filtering service can be sent a set of raw tags (e.g. an entire tag cloud), which will be processed and filtered by the service, and a *clean* set of tags will be returned back (details in Cantador et al. (2008)).
- **Sense Matching:** The TAGora Sense Repository (TSR) is a linked data enabled service endpoint that provides extensive metadata about tags and their possible senses. When queried with a tag, the TSR will attempt to find DBpedia.org URIs and Wordnet Synsets that correspond to the possible meanings of the tag. Since many of the tags used have multiple meanings (e.g. apple may refer to the fruit or the technology company), the TSR also provides additional metadata about DBPedia senses to assist in the disambiguation process (details are in Garcia-Silva et al. (2009)).
- **Profile Builder:** With the growth of Web2.0, it is becoming increasingly common for users to maintain a presence in more than one site. For example, one could be bookmarking pages in Delicious, uploading images in Flickr, listening to music in Last.fm, arranging social events with Facebook, etc.The Profile Builder service (currently not public) generate rich *Profiles of Interests* by bringing together and consolidating multiple folksonomy identities.

Cross-folksonomy data gathering and analysis was extensively used in building and deploying the **Live Social Semantics** application (LSS - Alani et al. (2009)) to identify various social connections between conference attendees. Such connections could be direct, based on their online social friendships, or indirect, such as those based on the similarity of their tag clouds Szomszor et al. (2008b), on their scientific communities of practices, and on their offline social contacts.

LSS uses the services above to support real-time social linking of individuals, used information gleaned from integrating data from various folksonomies. More detail about the services of LSS and the type of recommendation services it provides can be found in deliverable D4.5. An overview of the LSS application and its deployment is in deliverable D2.5. LSS is fully described in Alani et al. (2009).

Task 3.6 Collaborative tagging and emergent semantics:

Task 3.6a Improving Navigation for Images

In this section an overview of the initial results of the analysis of the Armin Link dataset will be presented, together with preliminary conclusions. The analysis performed consisted of two parts.

At first the Tag assignment distribution within in the Armin Linke dataset was examined, followed by a closer look at the three-mode network structure.

The measurements and examination of these measurements drew the following picture about the Armin Linke dataset. The constrains imposed on the underlying tagging system, such as batch tagging, anonymous users and limited set of resources had significant influence on the development of the folksonomy. Most changes occurred in the distribution of the users and the resources, but minor changes are also visible in the distribution of tags. In comparison to the common tagging system, for which the Delicious system was chosen as a representative example, the users form tiny cliques together with the assigned tag and the images used in the album the user used. These cliques are strongly interconnected through single tags and in a minor way through the multiple images as described by the Cliquishness and Connectivity measurements. The conclusion of these observations is that in contrast to usual tagging systems were a high interconnection among different tagged resources exists, the Armin Linke dataset is separable in many small topics which are interconnected lightly through tags. This can be interpreted as that the tagging process itself is less influenced by the popularity of the tags, in contrast to common tagging systems, and therefore supports a rather diverse usage of tags.

An investigative look into the actual tags, which are used within the Armin Linke dataset provides hints for the following conclusions. Most of the most popular tags are actually dates, and description of the location or event of the exhibition itself. Also tags in different languages, namely German, French and Greek, are dominant within the dataset and have often the same meaning once translated into English. A manual clustering and disambiguation of the most popular tags indicates that the set of tags can be significantly reduced and it is still to be determined how this densification would influence the network structure.

The analysis performed until now should be seen as a first step in the process to find an appropriate feedback for the users or to understand the true impact of the imposed constrains on the tagging process as present in the “Phenotypes/Limited Forms” installation.

The conclusions drawn out of this first look is that the current setup indeed supports a rather diverse tagging creation process with the drawback that the resulting network between users, tags and images is weakly interconnected. It is still to be examined if this is only the result of a slowed down growing process of the interconnections with the folksonomy, caused by external factors such as different languages, or a true result of the constrains.

The next steps in finding an appropriate feedback for the visitors of the exhibition lies in our opinion in the definition of a measure to determine the true diversity of an image based on the assigned tags. A first step to achieve this would be to automatically translate and maybe even disambiguate (perhaps using the web service developed by the team at the University of Southampton) the tags.

Task 3.6b Improving Automatic Classification of Music

Sony CSL also continued research in automatic tagging in the music domains, and came up with 2 main results. The first activity was the design and evaluation of efficient techniques for predicting automatically tag from the analysis of acoustic signals of polyphonic music. On the one hand, the team came up with a novel scheme to improve the acuracy of supervized classification techniques, using a kind of ensemble learning approach. on the other hand, the team performed an in-depth evaluation of the techniuqe on a large database of music (about 33.000 MP3s). The results of this study have been published in Pachet and Roy (2009). Many interesting insights were obtained, notably showing counter-intuitive effects (e.g. high-level descriptors are not necesarily more difficult to predict that low-level ones). Additionally, an original study concerning the possibility to predict Hit Songs using acoustic and manual metadata was performed, and published in Pachet and Roy (2008) This result was presented at the Ismir 2008 conference (Philadelphia), and received a great deal of attention from the music information retrieval community. However, the results were shown to be globally issuficient for completely automatic predictors. They were, therefore, not implemented in the Ikoru framework.

The other activity pursued by the music team addressed a novel way of looking at tags. The basic idea, coined “Description-Based Design”, consists in using tags not as description devices as in most tag-based applicatinos, but as tools to generate objects. A mechanism to pervert the basic Support Vector Machine (SVMs) structure was developped, in order to turn them into object generators. A pilot study was conducted in the domain of (musical) melody construction. In this study, tags describing melodies (such as tonal, jumpy, etc.) are used to construct new melodies according to user-chosen subjective dimensions. The study showed that the constructed melodies do optimize these subjective dimensions and was published in Pachet (2009).

2.3.3 Deviations and Corrective Actions

In case there is any deviation from the project objectives it must be described here by the task responsible:

PHYS-SAPIENZA: none

SONY-CSL: none

UNI KO-LD: none

UNIK: We have shifted our remaining resources from WP3 to WP4, based on the recommendations of the reviewers.

UNI-SOTON: Deliverable 3.4 has been wrongly named as Methods for tracking research topic emergence, whereas it should be Methods for tracking tag emergence.

2.3.4 Deliverables

For each deliverable please fill in all the missing information: actual delivery date, person months used.

Del. No.	Deliverable name	WP No.	Date due	Actual/ Forecast delivery date	Estimated indicative person-months	Used indicative person-months	Lead contractor
D3.4	Methods for tracking research topic emergence (Month 38).	3	Oct. 15th 2009	Sept. 15th 2009	1	1	**UNI-SOTON**

2.4 Workpackage 4 (WP4) - Modeling and simulations

2.4.1 Objectives

Following are the objectives of the research carried out during the third year of the project.

Task 4.1 Modeling:

This task aims at the partial reprodution of the structure of folksonomies. This problem might be tackled in different ways. One possibility is to implement so called stochastic models, by which the properties of folksonomies are recovered by means of recurrent random extraction of correlated tokens. This kind of methods have proven to be rather successful in reproducing the sought properties. Another possibility is to perform agent based simulations, where the folksonomy arises naturally from the interactions of individuals, in a way similar to language games. This last fascinating possibility is unfortunately far from the reality: there seems to be no clear interaction between users in a folksonomy, although this issue is still under debate.

Task 4.1.1- Stochastic Models

Even a partial reprodution of the properties of folksonomies is a formidable task. Last year we presented two successful models with the aim of getting the closest as possible to the real structure of folksonomies, starting from their global properties (tag frequency distributions, tag dictionary growth, etc) up to the subtle single post structure. The models we refer to are the *Epistemic Model* and the *Semantic Walker Model*. Despite their success, these two models were still at an embryonal stage. This year we purposed the objective to develop and further test these two models looking forward to their use as reference by the scientific community.

Task 4.1.2 - Language Games

See workpackage deviations in section 2.4.3.

Task 4.1.3 - Social Network Models

See workpackage deviations in section 2.4.3.

Task 4.2 Control :

Task 4.2.1 - Simulation and control on music and image sharing systems

Sony-CSL's contribution has been partially moved from WP4 to WP3. A motivation for this change can be found in the section *Deviations and Corrective Actions* of the first year's Periodic Activity Report.

The goal of this subtask is to provide feedback from the virtual tagging-application of Ikoru to the real world tagging-application. The final objective is to find evidence that a real-world tagging process, as the one realized in the experiments, truly supports the collection of many different, meaningful tags.

Task 4.2.2 - Ontology learning

Learning complete and correct ontologies from the huge and diverse set of tagging data is quite difficult as the data tends to be sparse and tag correlations may have different reasons like similarity or super and sub concept relations. However, when looking at more or leas restricted domains it is possible to achieve quite good classification results.

Task 4.2.3 - Simulation and control on bibliographic reference sharing system

The aim of the third project year was to bring back the models to the applications. For BibSonomy, this meant that we continued our research on semantic similarity measures, tag recommendations, user recommendations, spam detection, and the analysis of user behavior, and implemented corresponding features in BibSonomy. In Deliverable 4.5, we describe the research issues in more detail. In Deliverable 2.5, we explain how the results were implemented in BibSonomy.

Task 4.2.4 - Recommendations based on Network Analysis

Recommendation systems usually rely on the information they collect from monitoring their users activities within the system. Such systems are usually ignorant of what their users do outside of their systems. This task is to extend the work on recommendations by making use of user profiles that are generated by integrating their distributed folksonomy accounts. Once such integrations are in place, richer knowledge about what the users are interested in can be gathered and used for making cross-domain recommendations.

For comparing different kinds of recommenders, a unifying framework is necessary. With its new tag recommendation framework, BibSonomy allows now for such an evaluation in a real-life setting.

2.4.2 Progress

In this section the task responsibles must describe the progresses achieved during the third year of the project for each of the tasks described above.

Task 4.1 Modeling:

Task 4.1.1- Stochastic Models

Extending and Evaluating the Epistemic Model In the third year, we started a more fine-grained evaluation of the Epistemic Model that extends the evaluation already available in Dellschaft and Staab (2008). Furthermore, we extended the Epistemic Model so that it simulates complete postings instead of single tag assignments. The simulation of postings is done by drawing a given posting length from a pre-defined distribution that is taken from real tagging systems. Then the Epistemic Model ensures that in a posting every tag can occur at most once.

The objective of the evaluation was to quantitatively measure the distance between simulated tag frequencies and the real frequencies in selected streams from Delicious and Bibsonomy. Furthermore, the results should be compared to the respective results of the Yule-Simon Model with Memory Cattuto et al. (2007).

During the quantitative evaluation of the Epistemic Model, it became obvious that the simulated co-occurrence streams better reproduce the tag-frequency curves of streams from which spam postings were removed. This gave rise to the idea that the Epistemic Model may be used for computing features that help to classify users in tagging systems as spammers and non-spammers.

A summary of the quantitative evaluation of the Epistemic Model is available in D4.6. There, we also describe the model-based features which can be used for detecting spam in tagging systems.

Semantic Walker Model In the previous year, we proposed a model for social annotation systems. We showed that the process of social annotation can be seen as a collective but uncoordinated exploration of an underlying semantic space, pictured as a graph, through a series of random walks. This modeling framework reproduces several aspects, so far unexplained, of social annotation, among which the peculiar growth of the size of the vocabulary used by the community and its complex network structure that represents an externalization of semantic structures grounded in cognition and typically hard to access.

In the present year we carried out with an extensive evaluation of the model, mainly considering its robustness, as suggested by the reviewers of the project.

The work has been published on a prestigious journal (see Cattuto et al. (2009)). In particular the robustness analysis are provided as a supplementary information joined to the paper.

Robustness evaluation: semantic network topology We have considered networks with very different topologies in order to test the robustness of our approach with respect to the structure of the underlying network.

The Watts-Strogatz model has been put forward in Watts and Strogatz (1998) as an example of network with large transitivity and at the same time small-world properties, i.e. short distances between nodes. The construction procedure starts with a ring of N vertices in which each vertex is symmetrically connected to its $2m$ nearest neighbors (m vertices clockwise and counterclockwise). Then, for every vertex, each edge connected to a clockwise neighbor is rewired with probability p, and preserved with probability $1 - p$. The rewiring connects the edge's endpoint to a randomly chosen vertex, avoiding self-connections, and thus creating shortcuts between distant parts of the ring. For $1/N \ll p \ll 1$, a network with a large number of triangles (due to the initial ring structure) and small diameter (thanks to the shortcuts) is obtained. The degree distribution is homogeneous, i.e., peaked around its average value.

The random scale-free network obtained from the uncorrelated configuration model in Catanzaro et al. (2005), in contrast, has a broad degree distribution $P(k) \sim k^{-\gamma}$ (we have used $\gamma = 2.8$ and $\gamma = 2.3$) and a low clustering coefficient.

We have also considered the homogeneous Erdős-Rényi random graph, in which nodes are linked with a uniform probability. In this case, small diameter and small clustering coefficient are obtained, and the degree distribution is homogeneous. Moreover, a model of strongly clustered scale-free network in Dorogovtsev et al. (2001) has also been used.

Robustness evaluation: Distribution of post length We have performed numerical simulations on the model networks presented above, using random walks of fixed length, or of randomly chosen lengths extracted from a given distribution $P(l)$. For random walks of fixed length a sub-linear behavior is observed, followed by a saturation effect. For broadly distributed lengths, on the other hand, the sub-linear power-law-like behavior does not show any saturation, in agreement with some analytical insights.

Evaluation of the semantic walker model with real data-sets In order to test the Semantic Walker Model we considered the South Florida Free Association Norms database (http://w3.usf.edu/FreeAssociation/) as a proxy for a latent shared semantic graph. On this database we first performed, in collaboration with Alain Barrat from CNRS, Marseille, a statistical analysis of the word association graph (WAG) (directed and weighted) in order to identify the so-called Strongly Connected Component (SCC) (i.e., the set of all the words among which all the possible pairs are connected by at least a possible path), whose size turned out to be of 4845 words (see figures 2.8 and 2.9).

We then performed the random walks experiment using the word association graph as a latent semantic graph in order to check the robustness of the results against a realistic implementation of the latent semantic graph. Notice the the word association graph is directed and weighted. As already done with artificial graphs we compared the statistical properties of the co-occurrence graph constructed with random walk procedures performed on the word association graph with those of real co-occurrence networks of del.icio.us. Also in this case a strong similarity between the properties of the artificial and real co-occurrence graphs emerge. This proves that the idea of modeling the post production as a random walk process on a suitable latent semantic graph is reasonable and that the word association graphs turn out to be a good proxy for it.

Post structure analysis In order to better investigate the tagging user behavior, and inspired by our semantic walker model, we started a statistical analysis of the post structure. In particular, we considered several tag-streams, for instance tag co-occurrence streams (i.e. the series of posts containing a fixed tag). We then considered the global frequency f_i of each tag inside the stream. Now, restricting the focus on each single post, we can compute two very simple observables: the average frequency of tags in the post and the corresponding standard deviation. More precisely,

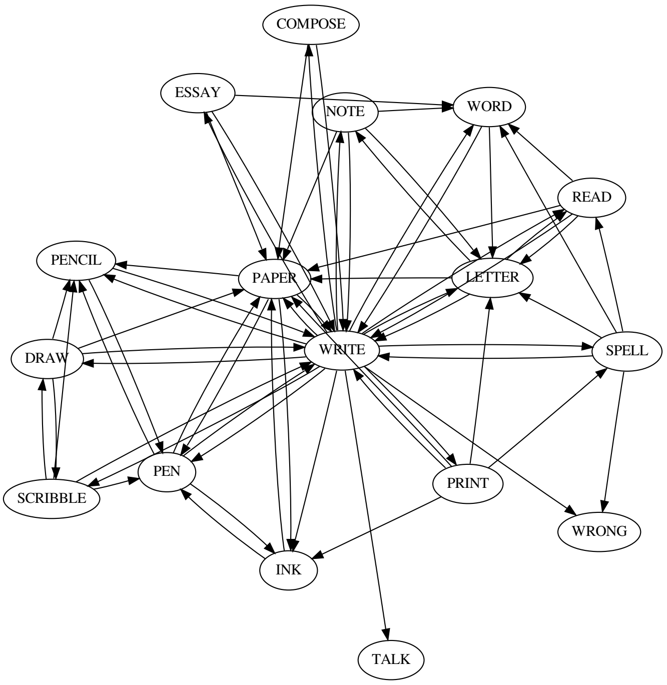


Figure 2.8: A small chunk of the whole word-association graph

for a post of length l:

$$p = \frac{1}{l} \sum_{j=1,l} f_j \tag{2.1}$$

$$\sigma_p = \sqrt{\frac{1}{l} \sum_{j=1,l} (f_j - p)^2} \tag{2.2}$$

where the sum is restricted to the tags in the post. The quantity of interest are now the averge of this quantities on all the posts with the same length l as a function of l.

In Fig. 2.10 we show the result of the analysis performed on the co-occurrence stream of the tag "phone" in the delicious dataset (panels a and b). As a null model we used a shuffled stream where correlation inside the post are obviously destoyed. The measures suggest that users tend to use more frequent tags for short posts, but when the post length increase they use less frequent (more specific) tags. We compare the measure with the post streams generated by the epistemic and the yule simon model, where no correlations are present in the post construction (panels c,d,e,f). Note how co-occurrence streams for more specific (less frequent) tags as "phone" or "gadget" behave more like the semantic walker model, where tags in a post are more correlated. On the other hand, co-occurrence streams of more generic tags (for instance "list") behave more like the epistemic or the Yule-Simon model.

Task 4.1.2 - Language Games

See workpackage deviations in section 2.4.3.

Task 4.1.3 - Social Network Models

See workpackage deviations in section 2.4.3.

Task 4.2 Control :

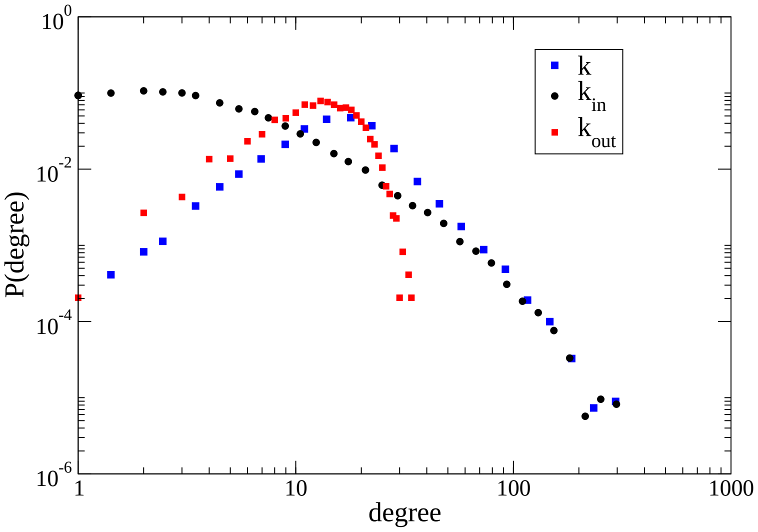


Figure 2.9: Degree distributions (k_{in}, k_{out} and k_{tot} for the word association graph

Task 4.2.1 - Simulation and control on music and image sharing systems

Armin Linke's "Phenotypes / Limited Forms" installation has been on display in 2008 at the Bienal de São Paulo in Brazil and the "Selective Knowledge" exhibition at the Institute for Contemporary Art and Thought (ITYS) in Athens, Greece and in 2009 at the "YOU_ser" exhibition at the Zentrum für Kunst und Medien (ZKM) in Karlsruhe, Germany and the "Concrete & Samples" exhibition in the Museum of Contemporary Art in Siegen, Germany. Visitors of the exhibition had the opportunity to tag such photos contributing to the Ikoru web-based system. Several observations have been made on the resulting data, such as the TAS element distribution and the study of the topological structure of the three-mode network constructed so, compared to a web-only built folksonomy such as Del.icio.us, adopted as a benchmark, as shown in table 2.4.2.

Characteristic path length	mean path length	Cliquishness	Connectivity
Armin Linke	3.5	0.95	0.14
Delicious	3.6	0.85	0.85

Table 2.3: Values of different network parameters for the Armin Linke and the Del.icio.us dataset

Task 4.2.2 - Ontology learning

Classifying Landmark Photos Creating general ontologies from the huge and diverse set of tags in folksonomies is a quite complicated task. It turns out that by concentrating on specific domains within a folksonomy one can achieve quite good classification results. For example, identifying landmarks is a popular problem. Our investigations show that by combining tagging data of Flickr photos with low-level features it is possible to identify with a higher probability than with other state-of-the-art research. Our approach can be generalized to other domains like car finding, mobile phone finding, etc.

Given the current widespread usage of digital photography, we observe users willing to share their

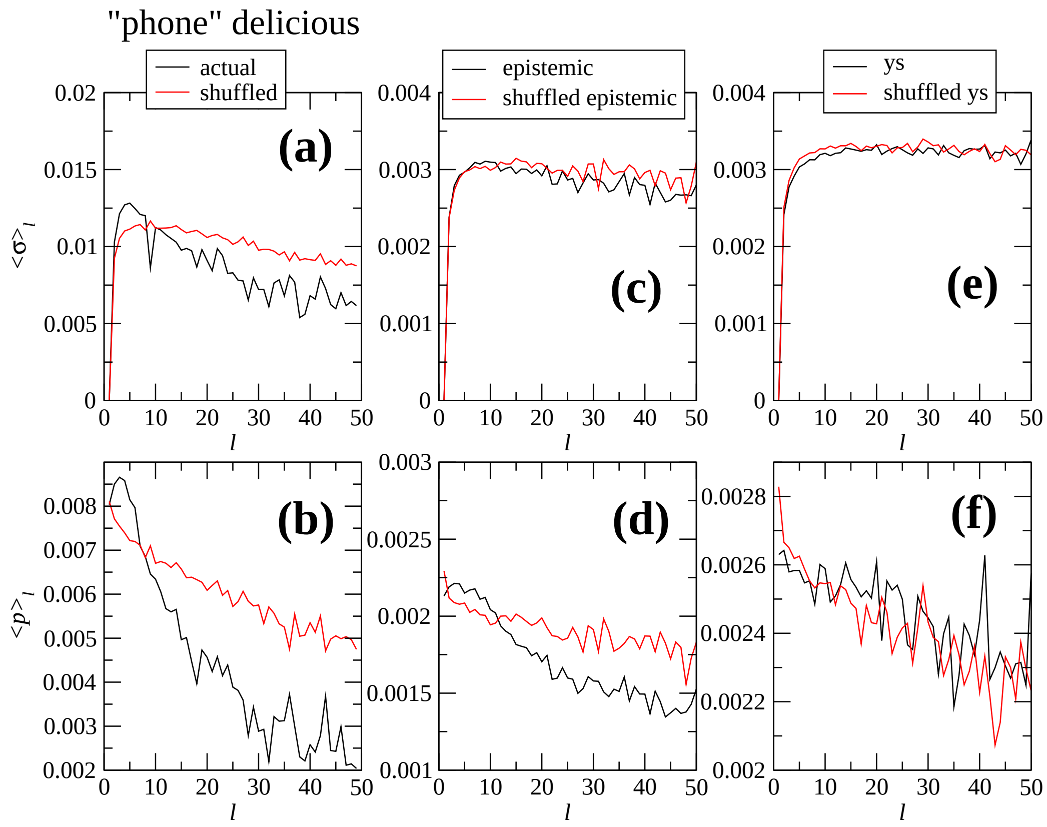


Figure 2.10: Statistical analysis of post structure for the "phone" co-occurrence tag stream in delicious, compared with several null models: shuffled stream, epistemic and yule-simon tagging models.

photos and experience within social platforms like Flickr[3]. As Flickr already contains billions of photos, the tasks of searching and navigating photos of interest become very difficult. To simplify these tasks, users adopted tagging, adding to each photo a set of freely chosen keywords. Still, simple tag matching does not give satisfactory results for particularly complex search tasks. One of such tasks is creating a photo summary of landmarks of a city, which is referred in literature as *landmark finding* problem. The World Explorer application Ahern et al. (2007) is the current state-of-art system which provides a landmark finding solution for Flickr. The system has a reasonable performance, but it only works with geo-tagged photos (supplied with geographical coordinates). The problem is that many interesting places around the world are still represented by photos without geo-tags and their landmarks cannot be found using World Explorer. The focus of our research is to exploit the tagging features and social Flickr groups to train a classifier with minimum efforts which can identify landmark photos.

Recognizing landmark in a photo is a hard task: First, content-based image analysis has very limited capabilities to solve this problem in general, given that photos are taken in different light and weather conditions, from different viewpoints and angles. Second, text-based or tag-based methods are much more appropriate for this task, but they do not have extra information if a tag represents a landmark or a family photo taken in a city. We propose to obtain this extra information from social groups in which users are involved in. Nowadays Flickr is enriched with specific photo groups related to landmarks, cars and other types of objects and themes, which can be used to distinguish the main topic of the photo.

[3] http://www.flickr.com/

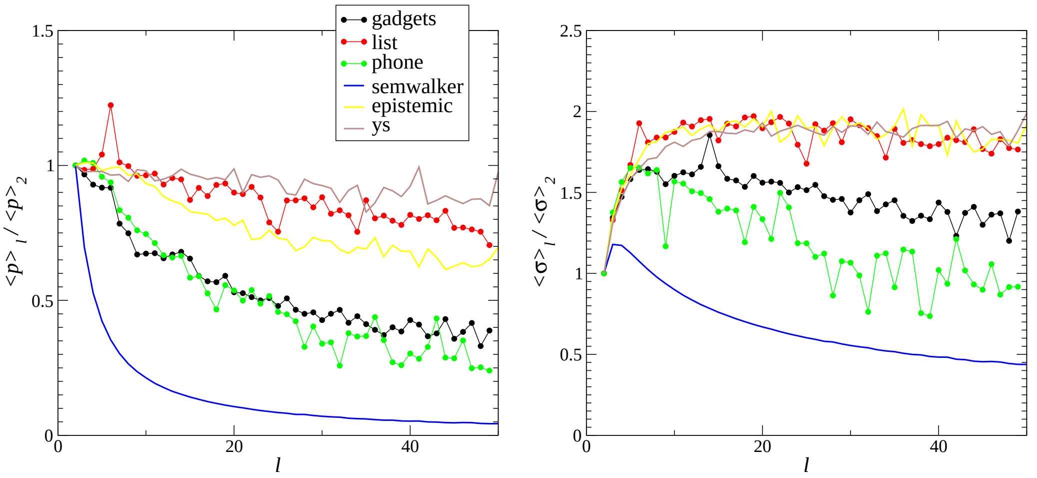


Figure 2.11: Statistical analysis of post structure for several co-occurrence tag streams in delicious, compared with synctetic streams, from semantic walker, epistemic and yule-simon tagging models. The quantity $\langle p \rangle$ and $\langle \sigma \rangle$ have been normalized to compare different tag streams.

Our proposed method contains two main parts. First, we exploit tags and social Flickr groups to train a classifier to identify landmark photos and tags. The method requires minimum human efforts, one only has to give links to relevant Flickr groups and the system automatically trains a classifier based on the data retrieved from Flickr groups. Second part of the method ranks all suggested relevant tags by their representativeness of a landmark. It is also possible to generalize our approach for other problems like car finding, mobile phone finding, etc. Although, due to high cost of user studies, we test the performance of our method for landmarks only. To the best of our knowledge, the proposed solution is the first one to solve landmark finding problem by exploiting tags and information from photo communities. Current method does not use low level image features or GPS-coordinates. Presented user study shows that our approach outperforms the state-of-the-art World Explorer.

User Study We performed a user study to evaluate the performance of our proposed method. We used the following data sets for evaluation:

Training Data (DS_{train}): The training dataset was used for training the landmark vs. non-landmark classifier. DS_{train} was constructed by downloading 430,282 photos from several Flickr groups, uploaded by 57,581 different users. For positive examples we manually picked few groups like “Landmarks”, “Landmarks around the world”, “City Landmarks”, etc. As negative examples we used groups like “Airplanes”, “Birds”, “Cars”, “Mobile Phones”, etc. The dataset thus created contains 14,729 positive examples (related to landmark groups) and 415,553 negative examples (related to general groups). None of these 430,282 photos was included in the test dataset. This is real-world data so “positive groups” might also contain some non-landmark photos and vice versa. However, no additional noise reduction technique has been applied.

Test Data (DS_{test}): This dataset consists of pictures corresponding to 50 cities (for which World Explorer Ahern et al. (2007) has at least 10 landmark tags), 60% European ones and the rest of 40% representing Asian, North-, South- American and Australian cities. We downloaded 4,000 to 5,000 photos/city, so that in total we gathered 232,265 photos, uploaded by 32,409 different users. Pictures from dataset DS_{test} were used for testing the classifier, after a model was learned based on DS_{train}.

The goal of our experiments is to evaluate the performance of the algorithm in finding city landmarks. We evaluate the accuracy of city landmark findings for the list of 50 different cities, included in the testing set DS_{test}, thus having in total 232,265 images at our disposal. The results of this analysis have been collected through a user survey. Additionally, with this user study we also compared our results against results produced by an existing system trying to solve the same problem, World Explorer Ahern et al. (2007). Since World Explorer uses as input for its algorithms Flickr pictures with GPS data – i.e. richer input data than we needed – our aim was to obtain at least comparable quality.

For the evaluation setup we recruited 20 volunteers among our colleagues. Each user was asked to evaluate two result sets for 10 randomly selected cities out of the set of 50, and the selection process picked each city so that by the end of the experiment it was evaluated by at least 4 users. Two photo summaries were mixed on a single screen, with one result set created using our algorithm and one coming from the World Explorer API. The users did not know which system produced which photo, as the photos from the two systems were randomly interleaved. Each photo was supplied with a title and a single landmark tag produced by either World Explorer or by our algorithm and used to retrieve this photo. A radio button was placed near each photo, where users could select between “landmark”, “non-landmark”, and “don’t know” options. The users were asked to judge if a photo is a landmark or not, in total producing between 400 and 500 judgments per user. The experiment took about 30 minutes per user.

Participants were instructed that a landmark photo must (1) contain a whole landmark or large part of it and (2) the landmark must be a main topic, not just a background for a person photo. Users were allowed to use photo title and tag as hints when they could not decide based on the picture only. The details of the results can be found in Abbasi et al. (2009).

Task 4.2.3 - Simulation and control on bibliographic reference sharing system

We have continued our research on measures for the semantic similarity of tags (in cooperation with PHYS-SAPIENZA), tag recommendations, the analysis of user behavior, and spam detection. The most promising measures and algorithms were implemented (in WP 2) into BibSonomy: the semantic similarity measures are now displaying related tags, similar tags, and related users; and there are frameworks for integrating online recommenders, for detecting spam, and for logging the user interaction. (For the recommender framework, only the design could be financed by TAGora; the implementation has been performed – during the life time of TAGora – within a national follow-up project.)

Task 4.2.4 - Recommendations based on Network Analysis

In the second year of the project we ran focusses on collecting user data, generating FOAF files automatically by mapping their tags to Wikipedia URIs, and running some experiments to test and evaluate these processes. The first experiment was aimed at understanding how similar user tag (Szomszor et al. (2008b)). The second experiment was to test the accuracy of building semantic profiles of users based on their tagging activities, and on the feasibility of using this information for generating news article recommendations, reaching average accuracy of class assignments was 69.9%, with 84.4% average accuracy of ontology instantiation (Cantador et al. (2008)). The third experiment we carried out aimed at expanding the above by investigating more closely how to semantically model user interests based on their distributed tagging activities (Szomszor et al. (2008a)).

In the third year of TAGora, the work above have matured into several freely accessible services (section 2.3), with large knowledge bases that contain all the necessary data for those services to run successfully.

Tag recommendations in BibSonomy. The BibSonomy recommender framework allows the integration and evaluation of different (semantic or ‘non-semantic’) recommender systems in BibSonomy. These recommender systems can be either installed locally or remotely (connected and queried via http), thus allowing other research teams to integrate their recommender systems and giving a broad base for evaluation. All incoming events and informations are logged for evaluation in a SQL database. The framework is described in more detail in Deliverable 2.5. The design of the framework has been performed within Tagora – its implementation was not possible with the Tagora budget any more. However, the framework could be implemented still within the life time of Tagora (financed by a national follow-up project). The recommendation framework is deployed in the ECML PKDD Discovery Challenge. See http://www.kde.cs.uni-kassel.de/ws/dc09/online for details.

Recommending Interests: In Szomszor et al. (2008a), we build a system that automatically generates a list of DBpedia URIs to represent interests a person might have by reasoning over their social tagging activity. With the LSS website, once a user has registered their social network systems accounts, any social tagging information from Delicious and Flickr is collected and converted to an RDF representation according to the TAGora tagging ontology[4]. For each of the user's tags, we use the TAGora Sense Repository (TSR) to lookup possible meanings of the tag, providing a mapping between tags and DBPedia URIs. Using the profile building algorithm described in D4.5, we were able to recommend to users a list of possible interests that they may want to expose to other conference participants. Users were allow to remove and add to the list of recommended interests before saving the results, thus providing us with feedback on the recommendation of these interests.

Recommending Tag Senses: Tags can often be associated to multiple senses (i.e. more than 1 DBpedia resource). To recommend the best sense for any specific tag, we build tools and services to compare the similarity (using a cosine measure) of the user's cooccurrence vector for any specific tag (i.e. all other tags that occur in the same post, and their frequencies) against the term frequencies associated with the possible DBpedia senses. If one of the similarity scores is above a threshold value, (0.3 in this case), we recommend that for that tag. If more than one (or zero) senses score above the threshold, we take a conservative approach and do not recommend any senses.

Recommending Conference Talks: This recommendation service was provided to conference attendees through the LSS application (see D4.5), where the person's social contacts in Facebook, Delicious, Flickr, and lastFM were crawled and overlayed with the list of authors of all papers presented at the conference. Users were recommended a talk if one of the paper authors is also their social contact in any of those social network systems.

Progress (Milestones)

M4.5 (Task 4.2) Preliminary control experiments (Month 32).

The control experiments performed are thoroughly described in the previous section. Substantial effort has been posed in the implementation of a feedback mechanism, where possible. Details can be found above.

2.4.3 Deviations and Corrective Actions

PHYS-SAPIENZA: In the modeling task, we ran into the following deviations.

[4] http://tagora.ecs.soton.ac.uk/schemas/tagging

Language Games (LG) is a general term indicating a set of models, based on the interaction of agents, aiming at reaching a sort of consensus. In folksonomies there seems to be practically no interaction between users so that LG models cannot have no rigorous justification. At the time of writing the proposal of this project, LG models seemed to be good candidates to partly explain the emerging folksonomy phenomenon. Unfortunately, as time elapsed, we realized that the sought strong collaborative character of folksonomy (intended as a collection of strongly interacting users) was no more justified. Therefore, we decided to divert our resources to the development of the much more promising stochastic models.

Social Network Models (SNM) are models involving the interaction of agents, who construct the folksonomy network. We realized, as for the LG models above, that the substantial absence of interaction between users left no justification for the study of SNM. The epistemic model already performs well in describing the folksonomy, and can be viewed as the result of a collection of independent agents. Therefore, we decided to focus on the improvement of the epistemic model.

UNI KO-LD: none

UNIK: none

UNI-SOTON: none

2.4.4 Deliverables and Milestones

For each deliverable and milestone please fill in all the missing information: actual delivery date, person months used.

Del. No.	Deliverable name	WP No.	Date due	Actual/ Forecast delivery date	Estimated indicative person-months	Used indicative person-months	Lead contractor
D4.5	(Task 4.2) Deployment of a semantic recommender (Month 38).	4	Oct. 15th 2009	Sept. 15th 2009	3	3	**UNI KO-LD**
D4.6	(Task 4.2) Report describing the results of the control (Month 38).	4	Oct. 15th 2009	Sept. 15th 2009	2	2	**PHYS-SAPIENZA (ALL)**

Mil. No.	Milestone name	WP No.	Date due	Actual/ Forecast delivery date	Lead contractor
M4.5	(Task 4.2) Preliminary control experiments performed (Month 32).	5	28 Feb 2009	28 Feb 2009	**PHYS-SAPIENZA**

2.5 Workpackage 5 (WP5) - Dissemination and exploitation

2.5.1 Objectives

The objectives for the third year are the same as for the previous year: to disseminate the research results, applications and strategies generated by the TAGora project within scientific and artistic communities, and also to communicate and apply them to a wide, general audience. Following are the objectives of the research carried out during the third year of the project.

Task 5.2 Dissemination strategies

During the third year of activity, the members of TAGora have focused their efforts in disseminating the outcomes of their research. As a result of this activity, the TAGora project is becoming a reference point for the scientific community and the general public interested in tagging. The main strategies for the dissemination of the project are based on the World Wide Web, through the TAGora website and its associated dynamic tools such as a blog and of course tagging. However, more diversified communication activities have been enacted outside the WWW. The knowledge generated within TAGora has also been featured in scientific papers and publications, general interest publications, poster presentations, conferences, talks and workshops.

Task 5.2.1 - Explicit dissemination activity

The dissemination activity, both online and offline, includes the following:

- Publication of research done within the scope of TAGora, in scientific journals and other written communication media, aiming to reach both the scientific community and the general public (refer to final PDK document for a full list of these publications);
- Presentation of research results obtained within the scope of TAGora, at different types of scientific and artistic events, such as conferences, talks, workshops, courses, demos and exhibitions (refer to final PDK document for a full list of these presentations);
- Maintenance of the TAGora website, which includes public documents and news;
- Constant and active feedback to web users, informing them about different news and events related to the project;
- Linkable content to establish contact with a broad online community;
- Release of research resources such as datasets and tools.
- Presentation and demonstration of applications on Conferences.

Task 5.2.2 - The role of applications developed in WP2

During the third year of the project, applications developed within TAGora should become crucial means to disseminate the research that is being made. These applications become the embodiment of the research itself, by reflecting the project's developments and new strategies.

Bibsonomy, developed by the University of Kassel, has aimed to increase its user base even further and provide new services and features coming directly from the consortium's research. Also they aimed to increase the usage of Bibsonomy by integrating it into third party library service.

The Live Social Semantics Experiment, designed by the University of Southampton and by the ISI Foundation (through the SocioPatterns.org project) was presented at the European Semantic Web

Conference 2009 and at the ACM Hypertext Conference 2009, with the objective to illustrate possibilities of utilising various TAGora technologies for the analysis of social connectivity of conference participants.

The tag based multi-source search engine MyTag, which was developed at the University Koblenz-Landau, aimed to be a demonstrator of TAGora technologies to a wider audience and to offer students at the University Koblenz-Landau an opportunity to actively work on TAGora related topics.

Finally, the team at Sony CSL team created the NoiseTube platform, an evolution of their Zexe.net system. The team has aimed to reach a broad audience by promoting this platform as novel means to support communities of concerned citizens.

Task 5.2.3 Contribution of Sony CSL

Every two years, Sony CSL organizes a symposium and open-house, which are a major opportunity to present and demonstrate our work to the scientific community. Two open-house events overlap with the TAGora project.

The laboratory has always sought to interact with the artistic community. These collaborations allow us to explore new interfaces or new usage of collaborative tagging and give us the opportunity to work with small but captivating communities.

Task 5.3 Training activities and outreach

The objectives of the training activities are to make the research done at TAGora available to scientific and general audiences, focusing on a hands-on approach. Direct training in courses and workshops can be possible, after the fruitful work during these months. Another important objective for the dissemination of TAGora is to reach an audience beyond the communities that are already interested or familiar with tagging. To achieve this, the team aims to take advantage of the significant impact already made on the artistic community, and take it even further by introducing tagging to new users and novel contexts.

2.5.2 Progress

In this section we describe the progress achieved during the third year of the project regarding dissemination strategies, explicit dissemination activities, the role of the applications and training activities and outreach.

Task 5.2 Dissemination strategies

Following the dissemination objectives, the members of TAGora have been publishing intensively, making presentations at conferences and workshops and giving lectures about the research and findings within the project. As expected, the applications and products played a significant role in dissemination.

The TAGora website was maintained during this third year, and its contents were enriched with news, publications, tagging datasets and tools to create and analyse folksonomies. The efforts to reach wider audiences and introduce them to tagging were successful, particularly in achieving the involvement of people which did not use tags as means of classification before.

Task 5.2.1 - Explicit dissemination activity

In the third year, findings about tagging which resulted from TAGora research have effectively reached diverse audiences, thus fulfilling the dissemination objectives. A number of articles and papers were published in scientific and non-scientific journals. These publications help to ensure that the results of the research done in TAGora are and remain widely available. Beside these publications findings were also presented at major conferences, talks and workshops throughout the world, providing means to achieve direct contact with interested audiences and peers. For example, the TAGora consortium was a gold supporter of the ACM Hypertext 2009 Conference which was held in Torino, Italy.

Regarding web-based dissemination, the TAGora website has been maintained and constantly enriched with new content. It is now an important point of reference for everyone interested in the study of tagging and folksonomies. New sections and resources have been added to it, continuing the path towards a portal-like concept. Simulators, tools to analyse folksonomies, datasets and new products delivered by the consortium, can now be downloaded directly from the website, accessed through offered APIs or by sending an e-mail to the owner due to licence agreement requirements. To improve the dissemination of the provided datasets further, the consortium followed the reviewers comment and offers all datasets in the universal RDF format. This content is enriched with tutorials explaining the functionality of the most popular social tagging websites available on the World Wide Web.

Scraping tools used to crawl data from websites like Flickr or Delicious were not published, since these tools are now out of data due to frequent changes in the publishing style and text formatting in these websites. Additionally, our legal contacts have advised us not to provide scraping tools for these websites since they might violate end-user agreements.

Task 5.2.2 - The role of applications developed in WP2

Applications developed within TAGora have played a very important role in the dissemination of the TAGora project. Currently all developed applications – BibSonomy, MyTag, Tagster, Live Social Semantics, Ikoru, Zexe and NoiseTube – are presented in separate sections of the TAGora portal. For applications which were made open source, such as Tagster and Ikoru, a link to the source code is provided.

Ikoru, the tag-based navigation application for images and music developed at SONY-CSL, was successfully put to use in the context of an installation by photographer and artist Armin Linke. The installation was exhibited in 2008 at the Zentrum für Kunst und Medien (ZKM) in Karlsruhe, Germany, the Bienal de São Paulo in São Paulo, Brazil and the “Selective Knowledge” exhibition in Athens, Greece. Now it is still on display in the Museum of Contemporary Art in Siegen, Germany. In this interactive installation each spectator can create a printed book with his or her own choice of Linke’s photographs and he or she is asked to label their book with a title. References to the TAGora project are included both in the printed books, the catalogue of the exhibition and in the space of the exhibition itself. However, to concentrate the efforts of the Sony CSL team on NoiseTube – the successor of Zexe.net – no further effort was invested to disseminate the Ikoru system beyond these exhibitions.

BibSonomy significantly increased its user base, thanks to its growing usefulness, since it is integrated in more and more library services and variety of features, which were implemented as a direct result of theoretical research. A reference to the TAGora project can be found in the about/projects section of the BibSonomy website.

The MyTag application, a cross folksonomy search and recommendation tool, achieved both of its dissemination goals: First, presenting MyTag on the demo and poster session of WWW 2008 and at the First Future Internet Symposium. After each of these events an increase of user activity on the platform was noticeable. Furthermore, MyTag attracted interested students who were thus exposed to TAGora technologies. This resulted in two bachelor dissertations for which students developed and described new MyTag functionality.

Live Social Semantics (LSS) is an application meant as a large dissemination window. It displays to the users how their tags have been collected from multiple sources, filtered, disambiguated, and merged. This is done by showing the users a list of their DBpedia URIs representing their interests, which have been inferred from their tagging activities. The application also shows the user the list of his/her friends who are attending the conference, as well as those with whom the user has had a face-to-face contact (using active RFID readings from the SocioPatterns.org platform). The LSS application was successfully deployed at two major conferences and has already gathered several hundred users.

During the first two years of TAGora, the Zexe.net platform, which explored new ways to use tagging – through mobile phones – in real-world situations and for the benefit of (offline) communities, was successfully deployed in cities both in Europe and South-America. However, starting from the third year new ideas about the application of tagging to environmental pollution led the team at Sony CSL to consolidate and extend Zexe.net to create a new platform called NoiseTube which focuses on the case of urban noise pollution.

NoiseTube is therefore a platform that enables a participatory approach to the monitoring and mapping of noise pollution by empowering individual and groups of citizens to tag an measure their personal exposure to noise using their mobile phone. Like Zexe.net the platforms aims to disseminate tagging concepts and technologies in a new domain (i.e. environmental monitoring) and among new audiences. While Zexe.net was geared to support small, targeted groups of citizens during short campaigns, the NoiseTube platform was designed and is being promoted to reach a much larger public of potential users. A reference to the TAGora project can be found on the NoiseTube website (http://www.noisetube.net) and in the papers that have been published.

To disseminate them easily among students and fellow researchers, the source code of the peer-to-peer tagging application Tagster and the Ikoru system was published under an open source license. By clicking on links provided on the TAGora website, anyone can download this source code and freely reuse or extend (parts of) it.

Task 5.2.3 Contribution of Sony CSL

The Sony CSL Open House, a bi-annual public symposium which was initially planned to take place in Paris in 2008, was replaced by an event in Tokyo for the occasion of the 20th anniversary of Sony CSL. Therefore an open, public presentation of the work done within the context of TAGora (Ikoru, Zexe.net and most importantly NoiseTube) was postponed until the Parisian Open House event now being planned for October 2009. At this event visitors will also be given the opportunity to experience the NoiseTube system first hand by making noise measurements and taggings in the street and seeing the result of their work on an interactive map afterwards.

The artistic installation of Armin Linke, titled "Phenotypes/Limited Forms", which is linked to the Ikoru system, continued to attract visitors at exhibitions during the 3rd year or TAGora. Currently it is still on display at an exhibition in the Museum of Contemporary Art in Siegen, Germany, which will remain open until 20 September, 2009.

Task 5.3 Training activities and outreach

Regarding training activities and outreach, the TAGora team has been involved in the following events:

- organiser - ECML PKDD Discovery Challenge 2008, September 15, Antwerp, Belgium, http://www.kde.cs.uni-kassel.de/ws/rsdc08
- organiser - Dagstuhl Seminar on "Social Web Communities", 21st - 26th September, 2008, Dagstuhl, Leibniz, Germany
- organiser - BOEMIE 2008 Workshop on Ontology Evolution and Multimedia Information Extraction, 2nd, December , 2008, Koblenz, Germany
- organiser - Workshop on Collaborative Construction, Management and Linking of Structured Knowledge (CK 2009), ISWC 2009, Washington DC, USA http://users.ecs.soton.ac.uk/gc3/

iswc-workshop

- gold supporter - 20th ACM Conference on Hypertext and Hypermedia,29th June-1st July 2009, Torino, Italy, http://www.ht2009.org
- organiser - 20th ACM Conference on Hypertext and Hypermedia,29th June-1st July 2009, Torino, Italy: Workshop on *Tagging Dynamics in Online Communities*
- organiser - ECML PKDD Discovery Challenge 2009, September 7, Bled, Slovenia, http://www.kde.cs.uni-kassel.de/ws/dc09

Milestones

M5.3 Sony CSL Open House
The bi-annual public symposium, initially planned to take place in Paris in 2008, was unfortunately cancelled. However, at the Open House event now being planned for October 2009 an public presentation of the work done within the context of TAGora (Ikoru, Zexe.net and most importantly NoiseTube) will take place.

2.5.3 Deviations and Corrective Actions

In case there are any deviations from the project objectives they must be described here by the task responsible.

Milestone M5.3 was cancelled (see above).

2.5.4 Deliverables and Milestones

For each deliverable and milestone please fill in all the missing information: actual delivery date, person months used.

Del. No.	Deliverable name	WP No.	Date due	Actual/ Forecast delivery date	Estimated indicative person-months	Used indicative person-months	Lead contractor
D5.5	(Task 5.3) Report on the impact, usability and user communities characterization of our web-based experiments and demos (Month 38).	5	Oct 15th 2009	Sept 15th 2009	2	2	**SONY-CSL**

Mil. No.	Milestone name	WP No.	Date due	Actual/ Forecast delivery date	Lead contractor
M5.3	The Sony CSL Open House 2008 biannual public symposium in Paris	5	28 Feb. 2009	Cancelled (see above)	**SONY-CSL**

2.6 Workpackage 6 (WP6) - Management

2.6.1 Objectives

Following are the objectives of the research carried out during the third year of the project.

Task 6.1. Management (Month 38).
The goals of this WP are: to co-ordinate the administrative and scientific work of the project; to ensure that the management plan is carried out; to monitor progress of the project and provide means to correct deviations from project goals; to ensure that the interface with the Commission runs smoothly; to continually evaluate the project's progress against project and WP objectives, quickly reporting any problems to management; to provide evaluation reports to the Commission as required.

2.6.2 Progress

Task 6.1 Management (Month 38.)
The Project Management was carried out by the project coordinator as well as by the Governing Board and node contractors. The project coordinator, Vittorio Loreto, has been and is responsible for the day-to-day co-ordination of the project and has been the main interface between the project and the European Commission. He allocated the financial contribution received from the Commission to the Contractors according to the "Programme of Activities" and the decisions taken by the Consortium. Moreover, the coordinator: (a) verified that the deadline, structure, and content of the deliverables prepared by the contractors are in line with what indicated in the contract, (b) addressed the Project Deliverables to the Commission, after prior validation by the Executive Committee.

The Governing Board (Vittorio Loreto for "Sapienza" University of Rome team, Luc Steels for Sony CSL team, Steffen Staab for the University of Koblenz-Landau team, Gerd Stumme for the University of Kassel team, Harith Alani for the University of Southampton team) was and is responsible for the political and strategical orientation of the project and for any important decision concerning the proper operation of the Consortium.

Contractors (Vittorio Loreto, PHYS-SAPIENZA; Luc Steels, SONY-CSL; Steffen Staab, UNI KO-LD; Gerd Stumme, UNIK; Harith Alani, UNI-SOTON) were and are responsible for: (a) coordinating the research, training and dissemination activities of their node on the basis of the contract and the decision taken by the Governing Board described above, (b) coordinate the preparation of the deliverables and reports for which are responsible, (c) produce a cost statement and an audit certificate every twelve months.

A detailed description of the more important management actions carried during the thirs year of the project are reported in section 3 of this document. A detailed description of knowledge management, training, and dissemination activities is reported in the Plan for using and Disseminating Knowledge (D6.4).

Progress (Milestones)

M6.5 Co-ordination and Management Meetings (Month 38).

Two Project meetings have been organized during the second year.

VI TAGora meeting Bagnovignoni (Italy), November 20-21 2008;

VII TAGora meeting Rome, May 7/8 2009;

Frequent contacts among participants were also maintained by e-mail, telephone, occasional visits, short and long term visits. Here is the list of the most important bilateral meetings:

- UNI-SOTON and UNI KO-LD organized a bilateral meeting in Koblenz, March 9-11 2009, focused on WP2.
- PHYS-SAPIENZA and UNIK organized a bilateral meeting in Rome, February 9-13th 2009, focused on WP2 an WP3.
- PHYS-SAPIENZA and UNIK organized a bilateral meeting in Kassel, March 30th - April 3rd 2009, focused on WP3 an WP4.
- PHYS-SAPIENZA and UNI-SOTON organized a bilateral meeting in Torino, June 27-30 2009, focused on WP3.
- PHYS-SAPIENZA and UNI KO-LD organized a bilateral meeting in Rome, July 13-14 2009, focused on WP4.
- UNIK and UNI KO-LD organized a bilateral meeting in Wurzburg, August 18-19 2009, focused on WP2-WP4.

2.6.3 Deviations and Corrective Actions

PHYS-SAPIENZA: none

2.6.4 Deliverables and Milestones

Del. No.	Deliverable name	WP No.	Date due	Actual/ Forecast delivery date	Estimated indicative person-months	Used indicative person-months	Lead contractor
D6.4	Yearly Management Report (Month 38).	6	Oct. 15th 2009	Sept. 15th 2009	4	4	**PHYS-SAPIENZA**

Mil. No.	Milestone name	WP No.	Date due	Actual/ Forecast delivery date	Lead contractor
M6.5	Co-ordination and Management Meetings (Month 38).	6	31 Aug 2009	31 Aug 2009	**PHYS-SAPIENZA**

Chapter 3

Consortium Management

3.1 Consortium Management

The Project Management was carried out by the project coordinator (Vittorio Loreto - PHYS-SAPIENZA -), and by the Governing Board. To foster collaborations among the partners, assure a proper evaluation of progresses and the identification of problems several Project meeting were organized and more specifically:

- **VI TAGora meeting** Bagnovignoni (Italy), November 20-21 2008;
- **VII TAGora meeting** Rome, May 7/8 2009;

Frequent contacts among participants were also maintained by e-mail, telephone, occasional visits, short and long term visits. Here is the list of the most important bilateral meetings:

- UNI-SOTON and UNI KO-LD organized a bilateral meeting in Koblenz, March 9-11 2009, focused on WP2.
- PHYS-SAPIENZA and UNIK organized a bilateral meeting in Rome, February 9-13th 2009, focused on WP2 an WP3.
- PHYS-SAPIENZA and UNIK organized a bilateral meeting in Kassel, March 30th - April 3rd 2009, focused on WP3 an WP4.
- PHYS-SAPIENZA and UNI-SOTON organized a bilateral meeting in Torino, June 27-30 2009, focused on WP3.
- PHYS-SAPIENZA and UNI KO-LD organized a bilateral meeting in Rome, July 13-14 2009, focused on WP4.
- UNIK and UNI KO-LD organized a bilateral meeting in Wurzburg, August 18-19 2009, focused on WP2-WP4.

3.2 Problems, deviations and corrective actions

PHYS-SAPIENZA: At the beginning of the second year PHYS-SAPIENZA experienced a serious problem related to the budget. When we received the second payment for the TAGora project an offsetting procedure towards the Sapienza University in Rome was carried out to the detriment of the TAGora Project. An offsetting of 72.750,27 Euro has been indeed done by the Commission in the second payment of the TAGora project. The reason for the offsetting is related to a very old project managed by a different department inside Sapienza University.

As of today the problem has not yet been solved due to the slow burocratic procedures among different faculties and departmemts of Sapienza University of Rome. The charge of the offsetting procedure has again been kept entirely on the PHYS-SAPIENZA team with the idea that a solution will come only after the end of the project. During the last year the Physics Dept. anticipated the needed budget to ensure the normal research activity of the project.

SONY-CSL: none

UNI KO-LD: none

UNIK: none

UNI-SOTON: none

3.3 Project Timetable and Status

Workpackages - Plan and Status Barchart (Third year)

Months	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38
WP1 - Emergent Metadata															D1.4 (Task 1.5)
WP2 - Applications															D2.4 (Task 2.2) D2.4 (Task 2.2) D2.5 (Task 2.1)
WP3 - Data Analysis and emergent properties															D3.4
WP4 - Modeling and simulations									M4.5 (Task 4.2)						D4.5 (Task 4.2) D4.6 (Task 4.2)
WP5 - Dissemination and exploitation									M5.3						D5.5 (Task 5.3)
WP6 - Management															D6.4 M6.5

Workpackages activities have started and are progressing as planned. Deliverables and Milestones have been achieved and delivered in time.

Chapter 4

Other issues

4.1 Co-operation with other projects of the Complex System Initiative

- Ciro Cattuto has been the General Co-Chair of Hypertext 2009: the 20th ACM conference on hypertext and hypermedia, June 29th to July 1st 2009, Turin, Italy. (`http://www.ht2009.org/index.php`).

- Andreas Hotho and Vittorio Loreto have been Chairs of the Track 2 on People, Resources, and Annotations of the Hypertext 2009: the 20th ACM conference on hypertext and hypermedia, June 29th to July 1st 2009, Turin, Italy. (`http://www.ht2009.org/index.php`).

- Andrea Capocci and Vittorio Loreto organized a workshop on *Tagging Dynamics in Online Communities* in the framework of Hypertext 2009, June 29th, Turin, Italy. All the TAGora teams presented their results at the workshop. (`http://www.tagora-project.eu/blog/2009/03/27/ht09-workshop-semiotic-dynami`

- Harith Alani, Vittorio Loreto, Steffen Staab, Gerd Stumme organized a Dagstuhl Seminar on *Social Web Communities*, September 21st to 26th 2009. (`http://www.dagstuhl.de/en/program/calendar/semhp/?semnr=08391`)

- Vittorio Loreto has been Session leader for the session on *Information and Communication Technologies* at the European Conference on Complex Systems ECCS08, Jerusalem, 10-19 September 2008. (`http://www.jeruccs08.org`)

Bibliography

Rabeeh Abbasi and Steffen Staab. RichVSM: enRiched Vector Space Models for Folksonomies. In *HYPERTEXT 2009, Proceedings of the 20th ACM Conference on Hypertext and Hypermedia*, Turin, Italy, 2009. ACM.

Rabeeh Abbasi, Sergey Chernov, Wolfgang Nejdl, Raluca Paiu, and Steffen Staab. Exploiting Flickr Tags and Groups for Finding Landmark Photos. In *ECIR'09: Proceedings of 31st European Conference on Information Retrieval / Advances in Information Retrieval*, volume 5478, pages 654–661, Toulouse, France, 2009. Springer Berlin / Heidelberg.

Shane Ahern, Mor Naaman, Rahul Nair, and Jeannie Hui-I Yang. World explorer: visualizing aggregate data from unstructured text in geo-referenced collections. In *JCDL '07: Proceedings of the 7th ACM/IEEE joint conference on Digital libraries*, pages 1–10, Canada, 2007. ACM. ISBN 978-1-59593-644-8. doi: http://doi.acm.org/10.1145/1255175.1255177.

Harith Alani, Martin Szomszor, Gianluca Correndo, Ciro Cattuto, Alain Barrat, and Wouter Van den Broeck. Live Social Semantics. In *Proceedings of the International Semantic Web Conference (ISWC)*, Westfields Conference Center near Washington, DC, 2009.

Per Bak, Kim Christensen, Leon Danon, and Tim Scanlon. Unified Scaling Law for Earthquakes. *Physical Review Letters*, 88(17):178501+, Apr 2002. doi: 10.1103/PhysRevLett.88.178501. URL http://dx.doi.org/10.1103/PhysRevLett.88.178501.

Andrea Baldassarri, Alain Barrat, Andrea Capocci, Harry Halpin, Ulrike Lehner, Jose Ramasco, Valentin Robu, and Dario Taraborelli. The berners-lee hypothesis: Power laws and group structure in flickr. In Harith Alani, Steffen Staab, and Gerd Stumme, editors, *Proceedings of the Dagstuhl Seminar on Social Web Communities*, 2008. URL http://www.dagstuhl.de/en/program/calendar/semhp/?semnr=08391.

Dominik Benz, Marko Grobelnik, Andreas Hotho, Robert Jäschke, Dunja Mladenic, Vito D. P. Servedio, Sergej Sizov, and Martin Szomszor. Analyzing Tag Semantics Across Collaborative Tagging Systems. In Harith Alani, Steffen Staab, and Gerd Stumme, editors, *Proceedings of the Dagstuhl Seminar on Social Web Communities*, 2008. URL http://www.dagstuhl.de/en/program/calendar/semhp/?semnr=08391.

Jeffrey A. Burke, Deborah Estrin, Mark Hansen, Andrew Parker, Nithya Ramanathan, Sasank Reddy, and Mani B. Srivastava. Participatory sensing. In *World Sensor Web Workshop (WSW'06) at ACM SenSys'06, October 31, 2006, Boulder, Colorado, USA*, October 2006. URL http://www.sensorplanet.org/wsw2006/6_Burke_wsw06_ucla_final.pdf.

Ivan Cantador, Martin Szomszor, Harith Alani, Miriam Fernandez, and Pablo Castells. Enriching Ontological User Profiles with Tagging History for Multi-Domain Recommendations. In *Proc. Workshop on Collective Semantics: Collective Intelligence and the Semantic Web (CISWeb 2008), in 5th ESWC, Tenerife, Spain*, 2008.

Andrea Capocci, Andrea Baldassarri, Vito D.P. Servedio, and Vittorio Loreto. Tag cloud alignment in Flickr social networks: a dynamical analysis. *submitted for publication*, 2009a.

Andrea Capocci, Andrea Baldassarri, Vito D.P. Servedio, and Vittorio Loreto. Statistical properties of inter-arrival times distribution in social tagging systems. In *HT '09: Proceedings of the 20th ACM conference on Hypertext and hypermedia*, pages 239–244, New York, NY, USA, 2009b. ACM. ISBN 978-1-60558-486-7. doi: http://doi.acm.org/10.1145/1557914.1557955.

M. Catanzaro, M. Boguñá, and R. Pastor-Satorras. Generation of uncorrelated random scale-free networks. *Phys. Rev. E*, 71:027103, 2005.

Ciro Cattuto, Vittorio Loreto, and Luciano Pietronero. Semiotic Dynamics and Collaborative Tagging. *Proceedings of the National Academy of Sciences (PNAS)*, 104:1461–1464, 2007.

Ciro Cattuto, Dominik Benz, Andreas Hotho, and Gerd Stumme. Semantic Analysis of Tag Similarity Measures in Collaborative Tagging Systems. In *Proceedings of the 3rd Workshop on Ontology Learning and Population (OLP3)*, pages 39–43, Patras, Greece, July 2008a. ISBN 978-960-89282-6-8. URL http://olp.dfki.de/olp3/. ISBN 978-960-89282-6-8.

Ciro Cattuto, Dominik Benz, Andreas Hotho, and Gerd Stumme. Semantic Grounding of Tag Relatedness in Social Bookmarking Systems. In Amit P. Sheth, Steffen Staab, Mike Dean, Massimo Paolucci, Diana Maynard, Timothy W. Finin, and Krishnaprasad Thirunarayan, editors, *The Semantic Web – ISWC 2008, Proc.Intl. Semantic Web Conference 2008*, volume 5318 of *LNCS*, pages 615–631, Heidelberg, 2008b. Springer. URL http://dx.doi.org/10.1007/978-3-540-88564-1_39.

Ciro Cattuto, Alain Barrat, Andrea Baldassarri, Gregory Schehr, and Vittorio Loreto. Collective dynamics of social annotation. *pnas*, 106(26):10511–10515, june 2009. URL http://www.pnas.org/content/106/26/10511.abstract.

Klaas Dellschaft and Steffen Staab. An Epistemic Dynamic Model for Tagging Systems. In *HYPERTEXT 2008, Proceedings of the 19th ACM Conference on Hypertext and Hypermedia*, 2008.

Klaas Dellschaft, Olaf Görlitz, and Martin Szomszor. Sense Aware Searching and Exploration with MyTag. In *Proceedings of ISWC09 Poster and Demo Session*, 2009.

S. N. Dorogovtsev, J. F. F. Mendes, and A. N. Samukhin. Size-dependent degree distribution of a scale-free growing network. *Physical Review E*, 63:062101, 2001.

Andres Garcia-Silva, Martin Szomszor, Harith Alani, and Oscar Corcho. Preliminary Results in Tag Disambiguation using DBpedia. In *In proceedings of the First International Workshop on Collective Knowledge Capturing and Representation (CKCaR'09), collocated with KCap 2009*, Redondo Beach, California, USA, 2009.

Daniel Grabs. Beschreibung und Evaluation des MyTag Merge Algorithmus. Master's thesis, Universität Koblenz-Landau, 2009.

Nicolas Maisonneuve, Matthias Stevens, Maria E. Niesen, and Luc Steels. Map and Measure Noise Pollution using Mobiles Phones. In Ioannis N. Athanasiadis, Pericles A. Mitkas, Andrea E. Rizzoli, and Jorge Marx Gómez, editors, *Proceedings of ITEE 2009 – Information Technology in Environmental Engineering 4th International Symposium*, pages 215–228. Springer Berlin Heidelberg, May 2009a. doi: 10.1007/978-3-540-88351-7_16.

Nicolas Maisonneuve, Matthias Stevens, Maria E. Niessen, Peter Hanappe, and Luc Steels. Citizen Noise Pollution Monitoring. In Soon Ae Chun, Rodrigo Sandoval, and Priscilla Regan, editors, *Proceedings of 10th Annual International Conference on Digital Government Research:*

Social Networks: Making Connections between Citizens, Data and Government, volume 390 of *ACM International Conference Proceeding Series*, pages 96–103. Digital Government Society of North America / ACM Press, May 2009b. URL http://portal.acm.org/citation.cfm?id=1556176.1556198.

Nicolas Maisonneuve, Matthias Stevens, and Luc Steels. Measure and Map Noise pollution with your phone. In Leah Buechley, Eric Paulos, Daniela Rosner, and Amanda Williams, editors, *DIY :: HCI – A Showcase of Methods, Communities and Values for Reuse and Customization*, pages 78–82, April 2009c. URL http://noisetube.net/publications/DIYforCHI2009.pdf. Proceedings of the DIY for CHI workshop held on April 5, 2009 at CHI 2009, the 27th Annual CHI Conference on Human Factors in Computing Systems (April 4-9, 2009 in Boston, MA, USA).

Benjamin Markines, Ciro Cattuto, Filippo Menczer, Dominik Benz, Andreas Hotho, and Gerd Stumme. Evaluating Similarity Measures for Emergent Semantics of Social Tagging. In *18th International World Wide Web Conference*, pages 641–641, April 2009. URL http://www2009.eprints.org/65/.

F. Pachet. Description-Based Design of Melodies. *Computer Music Journal*, 33(4), Winter 2009.

F. Pachet and P. Roy. Is Hit Song Science a Science?, 2008. Accepted to the International Symposium on Music Information Retrieval (ISMIR).

F. Pachet and P. Roy. Improving Multi-Label Analysis of Music Titles: a Large Scale Validation of the Correction Approach. *IEEE Transactions on Audio Speech and Language Processing*, 17 (2):335–343, 2009.

G. Pass, A. Chowdhury, and C. Torgeson. A picture of search. In *The First International Conference on Scalable Information Systems*, 2006.

Matthias Scharek. Optimierung von Suchmaschinen basierend auf dem Suchverhalten von Benutzern im Internet. Master's thesis, Universität Koblenz-Landau, 2009.

Rossano Schifanella, Alain Barrat, Ciro Cattuto, Benjamin Markines, and Filippo Menczer. Folks in folksonomies: Social link prediction from shared metadata, 2009. submitted for publication.

L. Steels and E. Tisselli. Social Tagging in Community Memories. In *Proceedings of the 2008 AAAI Spring Symposium*, Social Information Processing, Stanford University, California, USA, 2008.

Martin Szomszor, Harith Alani, Ivan Cantador, Kieron O'Hara, and Nigel Shadbolt. Semantic Modelling of User Interests based on Cross-Folksonomy Analysis. In *submitted to Int. Semantic Web Conf., Karlsruhe, Germany*, 2008a.

Martin Szomszor, Ivan Cantador, and Harith Alani. Correlating User Profiles from Multiple Folksonomies. In *Proc. Int. Conf. Hypertext (HT08), Pittsburgh, PA, USA*, 2008b.

Dario Taraborelli, Camille Roth, and Andrea Baldassarri. When groups blend with social networks: An analysis of community dynamics in flickr. *to be submitted for publication*, 2009.

D. J. Watts and S. H. Strogatz. Collective dynamics of "small-world" networks. *Nature*, 393:440, 1998.

Erica Westly. Citizen Science: How Smartphones Can Aid Scientific Research? *Popular Mechanics Magazine*, March 2009. URL http://www.popularmechanics.com/science/research/4308375.html.