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 01/06/2007 to 31/05/2008
Start date of project: June 1st, 2006
Due date of deliverable: July 15th, 2008
Distribution: Public

Date of preparation: 31/05/2008
Duration: 36 months
Actual submission date: June 20th, 2008
Status: Final

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

Contents

Chapter 1

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.

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 has dealt 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 data collection activity is a crucial activity of the project, aimed at acquiring and preprocessing relevant data, metadata and temporal dynamics, and at storing the acquired information in a form amenable for data analysis. The project web site has been extended with an additional page[1] that is specifically dedicated to the data collected and/or used by the project. On that web page, every data set is fully described, covering the type and quantity of data, format, and links to where the data can be downloaded from if the data is *public* or has been *anonymized*.

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.

Folksonomies have been known to exhibit striking statistical regularities and activity patterns. In particular, folksonomies are dynamical systems and each time a user tags a resource, the folksonomy grows: the whole tripartite network representing the folksonomy is an evolving graph with a complex dynamics.

In order to analyze the dynamical properties of the system, the first and most simple approach is to consider the stream view of folksonomy. In this case, the network structure is disregarded, or better, the network is projected in a zero dimensional space. It is thus important to introduce methods of analysis of macroscopic quantities associated to streams. These quantities are pretty simple to define, nevertheless some of them (eg. the dictionary growth) are hard to be explained.

At a further level of complexity stands the network analysis of folksonomies. The basic unit of information in a collaborative tagging system is a (tags, user, resource) triple, referred to as post. Further, each post can be split in multiple tag assignments (TAS), according on the number of tags in it. The global structure of a folksonomy is thus given by a tripartite graph with three kind of nodes (users, resources and tags) linked by hyperedges. In order to study this structure one can look global structure with specific projections of the tripartite graph. A very interesting network is the so-called co-occurrence network, based on post co-occurrence. Understanding the structure of these information networks is of paramount importance. First of all to tame their complexity, control them and devise new and more effective techno-social systems. On the other hand, since human users are attaching meaning to their annotations, the investigation of data structures, emerging out of this collective social dynamics, could possibly reveal something about the underlying cognitive and social mechanisms that create the collective annotation process. In collaborative tagging systems, correlations between tag occurrences are (at least partially) an externalization of the relations between the corresponding meaning and have been used to infer formal representations of knowledge from social annotations.

Modeling and simulations

The objectives of this research area are twofold:

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

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, predicting and controlling the emergent properties seen in the semiotic dynamics orchestrated in on-line communities. The modeling activity can be performed at different scales and looking at folksonomies from different perspectives.

First of all it is interesting to look at the stream view of folksonomies by introducing models for the tagging activity of an average user, in order to recover synthetic tag streams statistically similar to the one observed in real systems. An interesting point is how effectively modeling the background knowledge of the user. On the other hand, the exposition of the user to other users' activity influences his/her activity: consequently a special attention has to been paid at the role of the interface.

On the other hand it is important to introduce schemes and models to describe and understand the structure of the tag co-occurrence network. An important results of the last reporting period has been the introduction of a simple schematic modeling of an abstract semantic space, representing an hidden, shared concept map. This models allows the simulation of synthetic posts, whose corresponding tag co-occurrence network very closely mimics the statistics of real co-occurrence network. Moreover, the associated stream reproduces both the frequency rank distribution and the dictionary growth curves observed in real systems. The model allows the study of more refined statistical measures, more sensible to hidden semantic correlations in the tagging data.

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 are a mean to describe, search, and retrieve objects in an intuitive way, which constitutes an important factor of its success. TAGora is mantaining several 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 current managed systems are:

BibSonomy BibSonomy (`www.bibsonomy.org`, see Fig. 1.1), 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.

Ikoru Ikoru (`demo.ikoru.net`, see Fig. 1.2) is a prototypical system developed by Sony CSL 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.

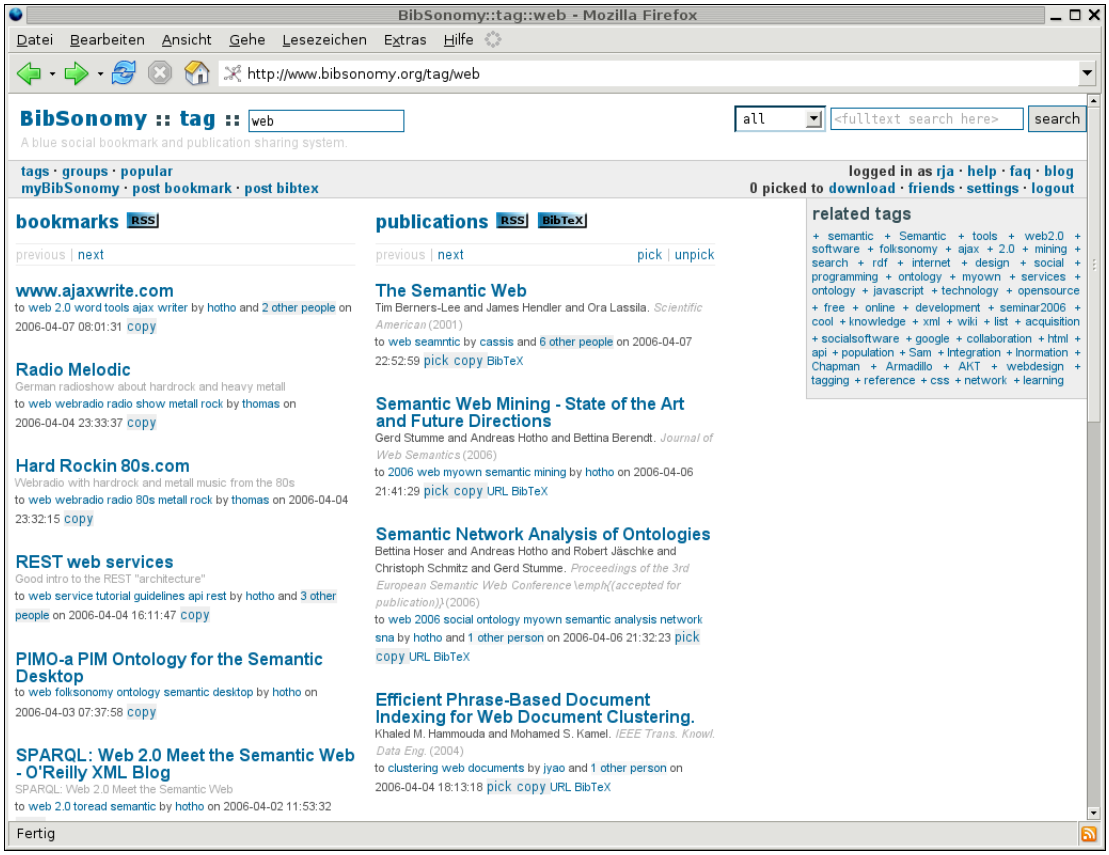

Figure 1.1: Screenshot of BibSonomy

Zexe Zexe project enables members of TAGora to study the dynamics of tagging in small-scale groups with shared interests. canal *MOTOBOY and GENEVE* accessible are the latest projects of the zexe.net initiative. Both of these projects make an intensive use of tagging. In canal*MOTOBOY, 15 motorcycle messengers in Sao Paulo, Brazil, use multimedia mobile phones to capture images and videos of their daily life. They use tags to describe these contents, which they publish on the web. The GENEVE*accessible project involves handicapped people in Geneva, Switzerland. They use multimedia phones equipped with GPS to create maps of their city's accessibility. They use tags to describe the images of obstacles they find in their way. By publishing these tagged and geo-referenced images on the web, they effectively build an intelligent, collaborative map which is immediately available to the public. Both groups have benefitted from the projects since they have allowed the participants to represent and communicate their particular issues.

Tagster Tagster is a peer-to-peer tagging application. Very much like Flickr, Del.icio.us, Bibsonomy etc. it allows to tag and share personal data. But instead of uploading the data to such an internet service, Tagster organizes and stores everything on the local computer. Therefore, unlike previous examples, Tagster can collect tagged resources of any format. It is based on a modular architecture, formerly known as the Semantic Exchange architecture (SEA). All tagged data are publicly available for the whole network and there is no mechanism to prevent the publication or to mark something as private data. The platform-independent version of Tagster can be downloaded via the project website. The knowledge of global statistics about tagging data extracted by Tagster is useful for different purposes. The next steps in the development of Tagster will include the collection of experience data from Tagster use, sophisticated means for recommending resources from the distributed peers, and additional wrappers for facilitating data collection from further seman-

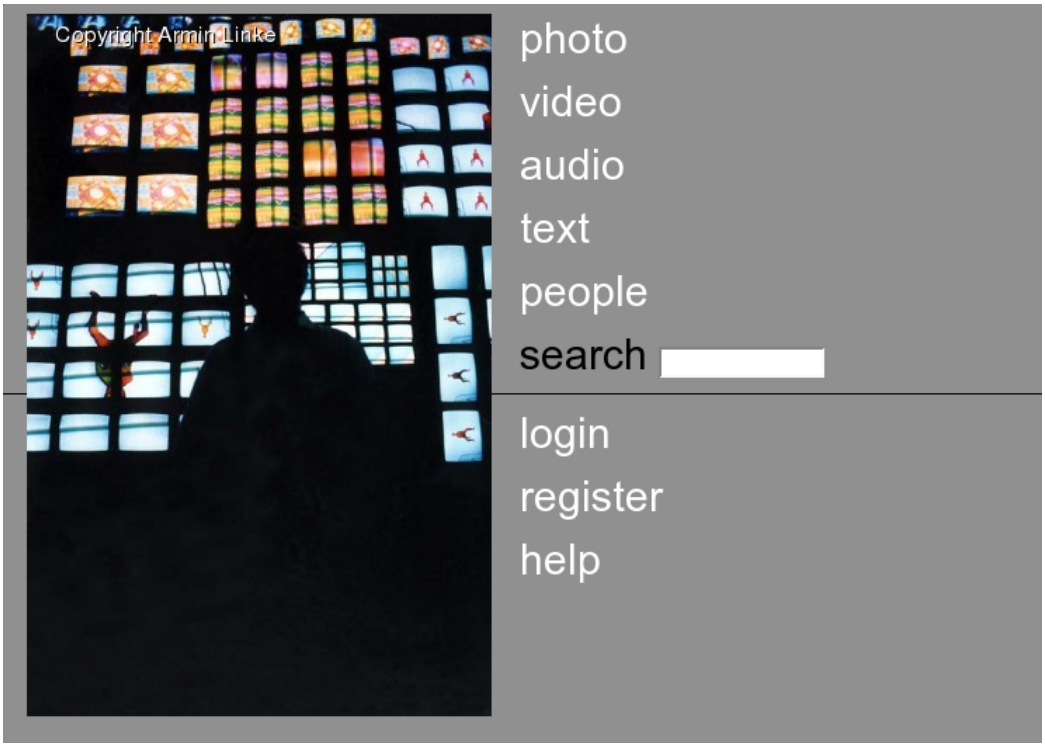

Figure 1.2: A screenshot of Ikoru from `http://demo.ikoru.net` .

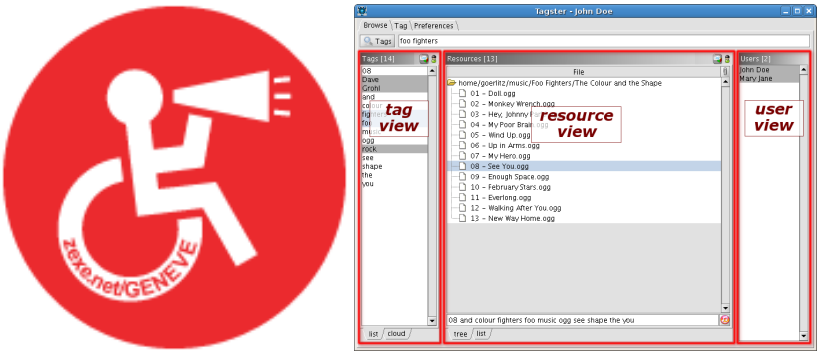

Figure 1.3: The Zexe logo (left) and A screenshot of Tagster (right).

tic and non-semantic sources. With such support Tagster may offer a viable open alternative to closed, centralized systems. The long-term objective is the efficient and effective infrastructure for decentralized, self-organizing Web 2.0 applications which allows for scalable sharing, annotation, searching, and browsing of relevant resources.

MyTag MyTag is a cross folksonomy search and recommendation tool set up by the University of Koblenz team `http://mytag.uni-koblenz.de/`. MyTag allows users to search accross folksonomy sites YouTube, Flickr and del.icio.us. Additionally, users can create a personal profile on MyTag that is used to tailor the search results to the users personal interests. This application was developed within the context of TAGora, and is being used to disseminate the project through the software implementation of theoretical research. MyTag is being used to track user interests and to gather search requests that can later be used for further analysis.

Figure 1.4: The MyTag logo

Results achieved so far

The main results achieved during the second reporting period include:

i) realization of three web-based applications: Zexe (zexe.net), Tagster and MyTag (http://mytag.uni-koblenz.de/);

ii) realization of the portal (Task 5.2) focused on collaborative social systems: to this end the TAGora web page has undergone a major restyling to accomodate information, tools and materials (e.g. data and simulators) addressed not only to experts from social sciences, information society, statistical physics but also to a general audience on the web.

iii) first data delivery through the project portal `http://www.tagora-project.eu/data/`;

iv) realization of a new web application for visualization and experiments on the tag co-occurrence network;

v) devising new concepts and tools for data analysis;

vi) implementation of the first control strategies on existing applications, in particular in Bibsonomy where an interface for spam management and the displaying of related tags have been set up;

vii) introduction of an epistemic dynamic model that can be used for simulating the assignment of tags to resources (Dellschaft and Staab, 2008);

vii) introduction of a first stochastic modeling scheme to mimic the construction of the tag co-occurrence network (Cattuto et al., 2008a);

viii) preparation of a review article on the *Statistical Physics of Social Dynamics*, now in press in Review of Modern Physics (Castellano et al., 2007);

Consortium and contact details

The project is coordinated by Vittorio Loreto (Physics Dept., *Sapienza* Università 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 2

Project objectives and major achievements during the reporting period

2.1 Project objectives

Social Tagging has gained a great impact in large-scale information systems. In applications like Flickr, Connotea, Citeulike, Delicious, etc. people no longer make passive use of online resources. Rather, they take on an active role and enrich resources with semantically meaningful information. Such information consists of terminology labels (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 aims 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 is being achieved through:

(a) a systematic and rigorous gathering of data 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 implemented in extensive numerical simulations.

TAGora aims 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.

2.2 Objective and main achievements of the reporting period

The main objectives of the second year of the project can be summarized as follows. First of all the Consortium was supposed to keep gathering data from already existing applications as well as the applications developed inside the project. At the same time the applications developed inside the project were supposed to grow, both in terms of number of users and in terms of new features. Given the availability of large datasets from different applications during the second year of the project the Consortium was more and more committed to the activity of extracting emergent features from the data and introduce modeling schemes to understand the underlying semiotic dynamics of online social communities. Finally a clear objective of the second year was to start the virtuos cycle the project aims at triggering between the data analysis and modeling activities and the applications. In particular the Consortium aimed at devising and implementing new features in the applications run by the project to set-up control strategies and enhance the navigation and the usability of the applications themselves. In summary the Consortium aimed at making further progresses along the main research directions of the project, namely:

- **infrastructures** developing tools and deploying data collection infrastructures (software, servers, network connectivity) for gathering data from collaborative tagging systems;
- **data analysis** devising methods and algorithms for analysing the raw data from the data-collection campaign;
- **modeling** developing theoretical constructions and models whose outcomes have to be compared with the experimental findings;
- **applications** developing and making publicly available innovative applications embodying novel navigation and control concepts;

Details of progresses achieved in these research lines are provided below. Here we only enphasize that significant progresses have been made in the three following areas:

A **Applications:** we realized several brand new web-based applications: (i) Zexe (zexe.net) devoted to a semiotic dynamics inside small communities interacting through state-of-the-art communications technologies; (ii) Tagster, one the first peer-to-peer tagging application; (iii) MyTag, one of the first cross-folknonomy application; (iv) TAGnet, a new web application for visualization and experiments on the tag co-occurrence network of a single user. We set up a portal focused on collaborative tagging systems through which, among the other things, the first public data delivery of TAGora has been realized `http://www.tagora-project.eu/data/`;

B **Data analysis:** new concepts and tools for data analysis have been explored. In particular different statistical measures of tag and resource similarity have been investigated (Cattuto et al., 2008b,c; Markines et al., 2008) and systematically characterized by means of semanting grounding in formal representations of knowledge. New analysis for the tag co-occurrence network has been performed and the emerging features submitted to the modeling activity. The analysis of tag streams have been also refined, focusing on correlations (Servedio et al., 2008), via standard two times correlators, as well as through an analysis of tag inter-arrival times. The last, in particular, allowed an interesting scaling analysis. The emerging picture shows that clustering of users' activity strongly affects the temporal evolution of collaborative tagging communities (Capocci et al., 2008).

C **Modeling activity:** As for the modeling activity an epistemic dynamic model that can be used for simulating the assignment of tags to resources (Dellschaft and Staab, 2008) has

been introduced that assumes two different main influences on the user during assigning tags: (1) the imitation of previous tag assignments made by other users and (2) the selection of tags from his background knowledge that he thinks are suitable for describing the content of the resource. On the other hand a first stochastic modeling scheme to mimic the construction of the tag co-occurrence network (Cattuto et al., 2008a) has been introduced. In this scheme it is shown that the process of social annotation can be seen as a collective exploration of a *semantic space*, modeled as a graph, through a series of random walks. Strikingly these simple assumptions reproduce several main aspects, so far unexplained, of social annotation, among which is the peculiar growth of the size of the vocabulary used by the community (Heaps, 1978) and its complex network structure (Cattuto et al., 2007c).

D Feedback to applications: implementation of the first control strategies on existing applications. In particular, Bibsonomy now features an interface for spam management and new navigation aids for browsing and moving among *similar* tags. As for Ikoru a cleanroom study on the inference of tags through the analysis of the audio signal has just been completed, even though the integration of this analysis technology in a web application still poses many challenges.

2.3 Summary of recommendations from previous review meeting and brief description of how they have been taken up by the consortium

In this section we summarize how the recommendations of the reviewers (reported in italic below) have been taken into account.

REVIEW REPORT N. 1 (covering period 01/06/2006-31/05/2007):

1 OVERALL ASSESSMENT OF THE PROJECT

The project has delivered according to the plans (WPs deliverables) and on several instances it is fair to say that the project has performed beyond expectations. The partners are already actively collaborating with a communication/integration effort genuinely driven by the scientific questions and the needs for different expertise.

The project concerns an area of increasing interest and it is going to face future competition both at the European and world-wide level. The consortium is however positioned to be one of the leading teams of the international research effort. The project is at the moment one of the state-of-the-art setters and is well positioned to produce some of the scientific breakthroughs in the field. On the other hand, in order to retain its position at the forefront of this research area, the consortium should start as soon as possible the proposed virtuous cycle data-gathering, characterization/modelling and application development. The reviewers believe that this cycle is the asset that might ultimately lead to the development of innovative applications and novel modelling and theoretical frameworks in the field. Another important "signature" feature of the project consists in providing data and tools that can be used by the community at large.

In summary, the project has proven to be well managed, with a coherent research program and the proper collaboration and expertise already in place. Tagora is positioned to deliver all the promised

results and very likely even more than what stated in the general research plan.

2 PROJECT ACHIEVEMENTS and FUTURE PLANS

WPs have delivered according to the plans and WP1 and WP2 are nicely ahead of schedule.

WP1 quickly gathered an impressive amount of data that are used to feed WP3 and WP4. The consortium has implemented a distributed plan of data gathering with different participants taking care of specific data collection efforts. It is possible to see that the project is ahead of schedule as consistent data sets have already been available for analysis in the WP3 and WP4.

The consortium was planning the implementation of software clients and hardware infrastructure for the collection of data. It appears that each specific data gathering project developed a specific client. While this might be the proper choice given the diversity of applications and data sets, it would be great if the project could provide a repository of clients and a shared infrastructure for data collection and publication.

While the public delivery of data is planned only for the end of the project, there is no reason to wait with this task. The publication of the data sets is expected to deliver a significant benefit to the larger research community, increase the visibility of the project and lead to new possibilities of external collaboration. (Preferably, both the data and the software code that was used to collect data should be released, although it may be the case with some of the sources that only the data or the software code can be made publicly available.) Therefore the reviewers would encourage immediate action on finding out the possible legal and ethical constraints of releasing data and software code publicly and proceed with releasing data and code as soon as possible, accompanied by a broad publicizing of this action and other support measures.

Description of how the suggestions/recommendations regarding WP1 has been taken into account by (see also point 6 of this report “CONCLUSION and SUMMARY of RECOMMENDATIONS”):

The TAGora consortium remains fully committed to the release of data and software whenever it is possible and beneficial to do so. To this end, the project web site has been extended with an additional page[1] that is specifically dedicated for the data collected and/or used by the project. On that web page, every data set is fully described, covering the type and quantity of data, format, and links to where the data can be downloaded from if the data is *public* or has been *anonymized*.

In a consultation with university legal advisers, we were advised *not* to publish controversially collected data, and any scraping software, until a few months before the end of the project, unless the data is fully and securely anonymized. This is to minimize the risk and effect of any possible objections from the data owners on the collection and/or republishing of their data. Such legal challenges, if risen, have the potential of forcing the project to stop using the data, and thus badly affecting its research agenda. Based on this information, for some of our data collection, the consortium will adhere to the original data-release date given in the project proposal, which is month 35. Nevertheless, the data collections that are affected by this issues will also be fully described on the web page mentioned above, to raise the community awareness of the existence of these data collections.

Currently, the data collections web page contains links for downloading the various datasets, such as Netflix, IMDB, Bibsonomy, MOTOBOY, UK and Last.fm music singles charts, IKoru, etc. The data that will be withheld for the time being include most data from Flickr, Last.fm, and del.icio.us. These datasets will be made available later in the project.

One of the recommendations given by the review committee last year was to gather user feedback to better evaluate some of the project results. We plan to launch a web site later this year where

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

users can view how their distributed tagging activities across various folksonomies (e.g. del.icio.us, Flickr, and Last.fm) have been compared and merged into user profiles of interests. Users can feedback on how their tags have been filtered, disambiguated, and merged, and on how their interests have been inferred from their tagging activities. This web site will build on the work described in deliverable D3.5.

Within the third project year, the BibSonomy logging mechanism will be extended, according to Task 2.1. The log data will allow us to study the various aspects of the user behavior in more detail by the given implicit user feedback (eg. click through data).

WP2 is in our opinion "overachieving" to say the least. In one year the consortium has already developed several relevant and fully functional applications, which is significant even considering that the project "hit the ground running", i.e. the applications or components thereof have already existed before the start of the project. The three applications have all reached a stage where they have been successfully deployed and providing valuable data to the project (Bibsonomy, Ikoru) or become ready for deployment (Tagster). In the case of Bibsonomy, the application has already attracted a user base that is certainly big enough for quantitative experimentation, e.g. bucket testing (different versions of the system shown to different groups of users for comparative evaluation).

For these reasons, in the case of Bibsonomy and Ikoru the reviewers would suggest to begin implementing the virtuous cycle set out as the primary goal of the project, i.e. and preparing support for user evaluations, implementing and experimenting with novel control mechanisms based on the findings of modelling and analysis. In the case of Tagster, a sufficiently large user community would need to be found and the remaining issues of distributed logging would need to be addressed so that this system can deliver data in a way that is comparable to the data provenient from the other systems. In particular, the data should allow the comparison of tagging behaviour in centralized systems vs. tagging behaviour in a peer-to-peer environment.

One of the primary uses of these applications is for controlled experiments and data gathering. However, in many cases it is possible to see that the developed applications might become very popular and being adopted by a large community of users (for instance Bibsonomy). In this perspective, the reviewers suggests an increased marketing effort and a more aggressive dissemination plan aimed at making these applications known outside the community of specialists in the field.

Description of how the suggestions/recommendations regarding WP2 has been taken into account by (see also point 6 of this report "CONCLUSION and SUMMARY of RECOMMENDATIONS"):

Concerning the cycle from models to control and back, we have implemented within BibSonomy an interface for spam management, which will be complemented in the third project year by a machine learning component for predicting spam users. The implementation of a tag recommender is also foreseen. Both tasks, spam detection and tag recommender, are also the subject of a dissemination effort, the Discovery Challenge of the ECML/PKDD conference 2008. This challenge is described in more detail in Deliverable 4.3. Further dissemination includes a system specific mailing list, cooperation with the Fraunhofer Institute for Autonomous Intelligent Systems and SAP Research, conference support for the Statphys23, ISWC+ASWC 2007, and ESWC 2008 conferences, references from several library catalogues, and publications about the system. Details are described in Deliverable 2.3.

The biggest challenge for Tagster is to acquire a sufficiently large user community for gathering tagging data that can be compared with the centralized systems. However, different aspects are influencing a software's usability and therefore the acceptance rate. For Tagster we have improved the navigation, e.g. by also including tag clouds, to resemble more closely what users know from

the centralized systems. Additionally, a novel mechanism for efficiently maintaining distributed statistics was developed and integrated. A problem, however, is still the use of the software behind firewalls, since the networking libraries used for implementing the distributed data management do not support firewall tunneling. However, a working solution is expected to be implemented in the beginning of the third year. So we will be able to continue disseminating the software. The data gathering mechanism is currently implemented in a basic fashion which requires some interaction of the user. In the third year this will also be extended to an automatic solution.

Concerning the implementation of the cycle from the system to models and back in Ikoru, we just finished the cleanroom study on the inference of tags through the analysis of the audio signal. The integration of this analysis technology into a tagging Web site still poses many challenges: it's computanionally intensive, the success rate of tag inference is low in most cases, and the tags on a public Web site are much more noisy and incomplete than those used in the study. At the current stage, a direct translation of this study into the Ikoru web site is not possible. However, to experiment with content-based analysis for audio, in the context of tagging, we integrated a fast kNN classification method for audio into Ikoru. The classification uses a subset of audio features that were used in the study mentioned above and is available as an interactive search tool in the Web interface. The practical experience that we gain from this experiment may lead to new ideas on how semantic inference can be transferred in a useful manner to collaboratively tagged data.

The recent developments of tagging sites for photos have put the entry level for new tagging sites very high. To compete with these sites and offer a satisactory and quality service to attract users is an uphill battle. Instead, we think it is in our interest to target Ikoru for small scale, specific projects. The artistic installation "Phenotypes/Limited Forms", which is on display in the ZKM museum in Germany, is a very particular application of Ikoru, yet, it has arguably gathered more data and reached more people than we would have through the marketing of the public Ikoru site. In addition, projects like these allow us to exploit new avenues for collaborative tagging, in particular how we can link this technology to day-to-day real-world situations, as can be seen in the Zexe.net projects.

WP3 shows a very good progress and greatly benefits from the earlier than expected input of WP1. The analysis and characterization are providing interesting results that can be used both to devise new data gathering project and in the development of theoretical models. It must be noted that the data analysis relies heavily on large scale network analysis. In this context we suggest the specific partners involved in this WP to consider increasing their presence at complex networks conferences and to proactively seeking for state-of-the-art analysis and characterization methods. In some of the presented results some of the recent developments and finding emerging in the area of complex networks characterization were missing (the consideration of proper null hypotheses, the analysis of mixing properties etc.). Further, we suggest that WP3 tackles more aggressively the issues related to the scalability of some of the methods used that while effective on small data sample can hardly implemented for large data sets and/or real-time contexts.

The project also faces intense competition in the semantic area, mostly due to the increased interest in the Semantic Web research field in analyzing and enriching folksonomies. Signs of this interest are the success of related workshops held at the last World Wide Web conference (WWW2007) and at the European Semantic Web Conference (ESWC2007) and the number of papers submitted to an upcoming issue of the Journal of Web Semantics on the topic of Semantic Web and Web 2.0.

In order to stay on top of the state-of-the-art, we advise the project to build on its unique combination of assets and start transferring the results of analysis into applications. This would allow the project to demonstrate its potential impact, in particular measuring the benefits for end-users in terms of improved browsing, more precise search and better recommendations or the removal

of spam. (Currently, the largest potential impact is expected to be in the area of recommender systems as exemplified by Ikoru.)
Evaluation through applications would also address the weaknesses in the evaluation of some of the methods, i.e. where quantitative evaluations are missing (e.g. clustering) and/or where the results cannot be judged by the authors themselves due to their subjective nature (e.g. ranking).

Description of how the suggestions/recommendations regarding WP3 has been taken into account by (see also point 6 of this report "CONCLUSION and SUMMARY of RECOMMENDATIONS"):

The science of complex networks is a very rapid evolving field. New methods of analysis have been constantly invented and new quantities have been studied. As for any field of science, correlation measures are crucial measures that are able to discern the trivial random essence of a network from the meaningful part due to cooperation and self organization. It is then crucial to refer to a null model of a network, against which the measured quantities on real networks may be compared. We faced this important problem from the outset by introducing the operation of network reshuffling. The reshuffling process destroys correlations present in the network and at the same time preserves the overall global features of the network. The possible presence of correlations (and of non trivial mixing properties) is then detected by simple comparison of the usual statistical indicators of the actual measured graph with the corresponding reshuffled one.

We used and adapted new method of analysis. To cite a couple, we performed the analysis of k-core component sizes of the folksnomies we crawled; the analysis of tag inter-arrivaltimes and tag-tag correlations, after considering folksonomies as temporal ordered streams, rather than as a network.

A significant issue is also the scalability of our methods. Until recently, RDF storage technology was limited to databases holding around 100Million triples, for example using 3store (http://sourceforge.net/projects/threestore/) or Sesame (http://www.openrdf.org/). We recently aquired Garlik's JXT, an RDF store capable of holding 60Billion triples (http://esw.w3.org/topic/LargeTripleStores). This signifcant increase in storage capability will provide us with the means to create large cross-folksonomy networks, linking the social networking activity and tagging history of many individuals across a vareity of sites. We are also developing novel tag disambiguation technology that will operate over a single user's data, rather than the current clustering techniques that require large sample datasets.

Tag classification with T-ORG is generally scalable to large systems as it has linear complexity and requires a very low amount of memory. The only bottleneck, however, is the interaction with the search engine (like Google) since queries are restricted to a certain number per seconds. This problem can be solved if we have special permission from the search engine to generate frequent queries and caching can also be applied to additionally speed up the process. There is also the idea to adapt T-ORG for client side application, such that search queries are send in a distributed fashion from the users computers.

Regarding the integration of the new algorithms into applications, we would like to point out that we have deployed FolkRank in BibSonomy. For a given tag, Folkrank generates not only a ranking of the resources (i.e., publications or bookmarks) most related to that tag, but also a ranking of the users that are most related to the given tag. This results in a "community" of interest around every tag in BibSonomy, together with a measure of participation. The community is displayed on the web page, and is returned in BuRST format. Moreover, we are currently shifting our focus on spam detection and tag recommendations in BibSonomy.[2] For example, our research on the notion of similarity in folksonomies (Cattuto et al., 2008b,c; Markines et al., 2008) led to a better navigation interface in BibSonomy: when browsing the page for a tag, the system now suggests a

[2]This is also part of our dissemination activities, see the ECML/PKDD 2008 Discovery Challenge at `http://www.kde.cs.uni-kassel.de/ws/rsdc08/`.

list of “related tags” (like most tagging systems), but also a list of “similar” tags, which can be used to overcome tag synonymy and morphological variations in tags. In general, the most promising approaches to tag recommendation and detection of non-social behavior will be implemented in BibSonomy during the third year of the project.

Southampton is planning on building an application over the summer of 2008 for recommendations that will use the output of T3.5 and T4.2.4. This application will also allow users to feedback on the results and thus help in evaluating our work.

WP4 is achieving good results and some simple stochastic models able to capture some of the basic features observed in the data characterization have been put forward.

Nevertheless, the reviewers' intuition is that improving on these basic models will require a deeper understanding of the user behaviour at the microscopic level. In particular, one would need to answer the question whether there is real imitation in folksonomy-based systems or the emerging properties are merely due to shared background knowledge. If there is real imitation, the question is also to what extent this is influenced by social factors, e.g. social networks and communities. Answering these questions is critical to applications, in particular finding out the extent to which folksonomies can be controlled and the best methods to achieve a certain goal (e.g. quicker convergence, less spam etc.)

As noted for WP3, what is missing is the impact of the WP4 results on the first two WPs. This is the feedback cycle in which new theoretical results trigger new data gathering endeavours and ultimately the development of new applications that on their turn poise new theoretical and analysis questions. The reviewers are aware that the project is just in its first year of life and are looking forward with confidence to the initiation of this virtuous cycle.

The resources allocated and expenditures are appropriate even if the consortium was not able to fully exploit the funding for personnel. We believe that this is a common problem during the initial year of every project as the finding and hiring of post docs might require several months. The consortium has however invested in computational resources and we are confident that the budget will be appropriately utilized in the next two years.

Description of how the suggestions/recommendations regarding WP4 has been taken into account by (see also point 6 of this report "CONCLUSION and SUMMARY of RECOMMENDATIONS"):

Since the beginning of the project, we posed ourselves the question of how much user co-operation is inside folksonomies and to wich extent the nice and interesting properties of them, rather rely on shared background knowledge. This question is hard to answer. We developed and are continuosly developing models to help unravel this point. The case of del.icio.us, in which users are exposed to tag suggestions by the system, is particularly complicated from this point of view and poses more challenges.

3 CONSORTIUM PARTNERSHIP

Despite this is the first year of the project, the consortium appears as a cohesive group of synergetic partners. The cooperation of the partners in the reporting period is good and there is no sign of an underperforming partner. The advantage of a small consortium is in fact a mutual dependence: the underperformance of a single partner would put the success of the entire project at risk.

Cooperation among partners will need to intensify in the coming period, as the interdependence of the tasks is increasing.

PHYS-SAPIENZA the level of cooperation among the different teams has increased in the second reporting period, through concrete exchanges of researchers but mainly through a process of convergence of the research topics addressed by the different teams. This is happening despite the different skills and backgrounds of the teams involved in the project. Examples in this direction are: PHYS-SAPIENZA and UNIKO-LD both worked on modeling streams of tags though from different perspectives; UNI-SOTON and UNIKO-LD both worked on integrating folksonomies; PHYS-SAPIENZA and UNIK worked intensively together to include new features in Bibsonomy (spam detection, detection of similar tags, etc.); PHYS-SAPIENZA and UNIKO-LD both worked on the usage of background knowledge in data analysis; PHYS-SAPIENZA and SONY-CSL collaborated in the analysis of data coming from zexe.net. New collaborations on different research topics are always growing also stimulated by bilateral meeting and the general TAGora metings.

4 PROJECT MANAGEMENT AND CO-ORDINATION

The management of the project is commended on their outstanding work in organizing the collaboration among a diverse set of partners with very different skills and backgrounds (in particular, participants working in the areas of Complex Systems and Computer Science). Besides the individual efforts of the partners this has resulted in a very successful kick-off of the project with a number of significant results achieved in the first year. Collaboration among partners is currently most visible in the area of data collection and in the creation of a White Paper outlining the challenges facing the project and the research field at large. On the other hand, we expect that more efforts will need to be spent in the future in integrating results from various methods of analysis and modelling. It will also be the task of the coordinator to coordinate the validation of results through applications and in general avoid the duplication of efforts in application development.

PHYS-SAPIENZA see above for a discussion about the level of cooperation reached in the project.

The reviewers also appreciate the swift actions of the management in adapting the planning of the project to external circumstances (e.g. difficulties in hiring) and taking corrective action when realizing that a better division of tasks is possible among WP3 and WP4.

The reviewers would suggest some changes in the way results are communicated to the PO and the reviewers. With respect to the deliverables, the reviewers' impression is that the documents could be more focused. In particular, the documents should not contain material directly copied from scientific papers, but instead summarize the findings and illustrate them to the level that is necessary for their understanding in the context of the project. (Scientific papers containing the details should be made available separately to the PO and the reviewers.) The context of the project should also be the primary basis when writing the introductions and conclusions of all reports, i.e. what are the questions that the project requires to answer with respect to the deliverable, to what extent these questions have been answered and how the results will be applied.

Deliverables length has been substantially reduced with respect to the report of last year. Content of deliverables in now much more fucused on main achievements and the overlap among the different deliverables and reports have been avoided. We now refer for details to the list of scientific papers that is available separately.

With respect to the special case of the Periodic Activity Report, the reviewers would suggest to significantly cut back the size of this report in the coming periods by leaving out the detailed description of the results. (These details are found in the other deliverables and therefore it is not necessary to repeat them.) The particular PA submitted for this period contains even more details than the individual deliverables themselves, which should certainly not be the case.

We fully accept the suggestion of cutting back the size of the present report and demand the detailed explanation of results to the deliverables and the attached list of scientific papers.

5 USE AND DISSEMINATION OF KNOWLEDGE

The dissemination and outreach activities are excellent. The web site of the project is sleek and informative. The leaflet on the project is extremely effective, as well as the white paper (deliverable). The project is organizing a major conference that is attracting considerable attention also outside the community of experts. Dissemination through artistic means (tagging of real world object) organized by Sony CSL is interesting in that it can reach a potential consumer audience beyond the usual circles of academia and industry. The reviewers look forward to extensions of these experimentations. Dissemination through popular science media (by PHYS-SAPIENZA) is also an excellent way to build a name for the project.

On the negative side, the project name and logo is not apparent on the main visible outcomes of the project, the Bibsonomy and Ikoru demonstrators. The project name and logo, along with a link to the project website should be apparent on the welcome page of these websites, allowing visitors access to the project website.

The Tagora logo together with a link to the project website has been addded to the projects page of BibSonomy. We decided against putting the logo on the welcome page, since this would set the signal that BibSonomy is a temporary activity only; and would increase the risk that users do not contribute their content any longer.
The TAGora logo has been inserted in the about Ikoru's page.

The management of the project is also strongly encouraged to communicate -without delays- any breakthroughs, major results and successful dissemination actions (such as the publishing of high visibility articles in traditional/popular media) as these events occur through the year. This would allow the reviewers to keep up-to-date with the project and enable the European Commission to take possible secondary dissemination actions such as the publishing of press releases.

We gladly accept this suggestion and we commit ourselves to timely disseminate major results to the Commission and to the reviewers.

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

The work plan seems coherent and progresses along the path outlined in the technical annex. Several WPs are ahead of schedule and performed above expectations. There are all the requisites to ensure that the project will deliver all the deliverables listed for the second part of the project. Some recommendations:

-WP1 should start considering as soon as possible the issue of data publication and availability. Data collection will most likely also need to include the explicit gathering of user feedback through end-user studies: there are several areas of proposed improvement where the progress cannot be measured without involving end users.

-WP2: The consortium should provide extra efforts to devise dissemination strategies for the "applications" developed in the project (see for instance Bibsonomy). Application development needs to be coordinated so that the three systems (Bibsonomy, Ikoru, and Tagster) allow evaluating different

aspects of the project.

-WP3 and WP4 should tackle more aggressively the issues related to the scalability of some of the methods used that while effective on small data sample can hardly implemented for large data sets and/or real-time contexts. In the future the network analysis should benefit of more contact with the progresses obtained in other areas, especially the foundational theory of networks. The existing results should be integrated and applied to particular tasks, e.g. recommending images, improving navigation through ontology learning etc.

- While the project is well performing in terms of outreach and dissemination, the reviewer sees extra potential that should be exploited. Timely communication with the EC for extra dissemination and more exposure of the "Tagora" logo on the developed applications are suggested. Similarly, a more aggressive outreach plan in related areas (network science, complex systems) could be beneficial.

Chapter 3

Workpackage progress of the period

Following are the objectives that were planned for all the workpackages for the second year of activity of the project. For the tasks started in the first year see the attached file PA1.

3.1 Workpackage 1 (WP1) - Emergent Metadata

3.1.1 Objectives

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

The objectives of this WP remain unchanged, which is to collect the required data to carry out all the research and investigations detailed in the other WPs, and to coordinate the efforts of collecting and storing the data whenever necessary. 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. del.icio.us 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 have now started to emphasize and encourage data sharing whenever possible. As explained earlier in this document, a new web page has been added to the TAGora project web site for describing our data collections and for pointing the public to where some of the data sets can be obtained from.

The following describes the objectives of each task in this WP. An update on data collection activities with respect to the four tasks of WP1 will be given in 3.1.2.

Task 1.1 Data from collaborative tagging (folksonomies)

In view of the publication of crawled data on the web site of our project, data acquired so far need to be validated and checked for consistency. Objective of this task is to perform this validation. In connection with Task 4.2.3, methods to detect and isolate spam activity will be devised and tested. Results on this point will be presented in deliverable 4.2. We plan to continue the monitoring of newcoming users in del.icio.us and Flickr, in view of next future massive crawl.

Task 1.2 Data collection from the bibliographic reference sharing system BibSonomy

The data of the BibSonomy system is going to be made public for scientific purposes. Particular attention is being paid in the anonymization procedure, after which it will not be possible to infer

the identity of users.

Data Collection from Tagster Tagster (cf. D2.3) is a decentralized collaborative tagging application. It provides the same functionalities as found in common centralized folksonomy systems like Bibsonomy. Users can tag multimedia resources with arbitrary text labels (tags) and browse the peer-to-peer network via the tags, users, or resources. Although the user interaction is basically the same as in the centralized case we expect to find different characteristics in the tagging data. Therefore, all user's complete tagging data is gathered for further analysis. Additionally, the Tagster clients also log information about the user searches.

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

The dataset from the canal*MOTOBOY project, which involves a small-scale community using tags to represent and communicate their daily life experiences has been made available to the TAGora consortium. In canal*MOTOBOY, 15 motorcycle messengers in Sao Paulo Brazil transmit tagged images, videos and audio clips directly from their mobile phones to a web page. The dataset, which includes 13 months of activity, can be used to study the dynamics of tagging of a small, densely-connected group. It contains 8.079 tag assignments, 7.975 resources, 712 tags and 15 users.

Phenotypes/Limited Forms is an art installation that is built on top of the Ikoru system. The installation has been on display at the Zentrum fur Kunst und Medien (ZKM) in Karlsruhe, Germany, and at the Selective Knowledge exhibition in Athens, Greece. We collected data about 8000 users, 1000 photos, 8000 tags, and 70000 tag assignments. The data gathering started in November 2007. The photos are copyrighted, but the tag assignments are available.

Task 1.4 Collecting data from online recommendation systems

As part of WP4, we plan to implement a system that uses information from folksonomies to generate personalized recommendations. In year 1 of TAGora we collected and studied the Netflix data for making recommendations based on information gathered from IMDB to investigate the benefits of combining such resources to improve recommendations. In the second year of the project we shifted our recommendation analysis work to more folksonomic resources (i.e. community and tagging based) such as del.icio.us, Flickr, and Last.fm. The objective here is to gather information from these folksonomies to allow us to generate cross domain, personalized, recommendations.

3.1.2 Progress

Task 1.1 Data from collaborative tagging (folksonomies)

Data collected last year on del.icio.us and Flickr folksonomies were tested for consistency. As a result, few records were removed and many others were retrieved again in order to recollect missing data. The set of del.icio.us and Flickr users is still constantly being monitored and newcoming users id's are added to our database in view of another more extensive future data crawl. We proposed spam detection methods based on topological, semantic and user profile data, testing them with success on the Bibsonomy platform. More on this point can be found in task 4.2.3.

Task 1.2 Data collection from the bibliographic reference sharing system BibSonomy

BibSonomy dumps can be downloaded for scientific purposes from the BibSonomy webpage, see `http://www.bibsonomy.org/faq#faq-dataset-1`, after having signed a license agreement. The delivery process is stable; the only modification we made is that the dumps are now generated every three months (instead of every half year).

Data Collection from Tagster Tagster stores two different sets of user data which will be useful for later analysis. The first one is the tagging data itself which is kept in a local RDF database. The second one is a log file containing all data search queries of the user. Towards the end of the second year Tagster was distributed within the ISWeb research group at Koblenz extended real-world testing before releasing it to the general public. The data gathered within that user group can now be used for first investigations on the tagging behavior in a decentralized scenario.

Due to the short period of data gathering we just obtained a small dataset. With the extension of the user base we should be able to collect a more representative dataset. However, that will also require a sophisticated data gathering mechanism to be implemented as client may be offline for a longer time period. So far we can still directly access the stored data of all participating users.

The local data of all users is merged into one big data set with following structure: In the tagging data set we store a timestamp, userID, resourceID, and tag for each tag assignment. The user and resource IDs are anonymised. The query log contains for each item query a timestamp, the queried item set and the associated item type. For the latter one there is no anonymisation necessary.

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

The canal*MOTOBOY dataset represents the tagging activity of 16 users during 13 months. In this period, the participants of the project created 712 different tags, 7.975 messages (resources, which can include more than one multimedia element) and 8.079 tag assignments. The dataset is already available to the TAGora consortium, and can be downloaded at `http://www.csl.sony.fr/~tisselli/zexe/zexe_motoboy.csv` In this dataset, which is stored as a comma-separated file, we include messageID, userID, date and tag for each tag assigment.

Task 1.4 Collecting data from online recommendation systems

Additional data has been collected this year to support research on cross-folksonomy analysis (Task 3.5) which will feed into the work on recommendation systems. For this work, we needed to gather data about users and their tagging activities from multiple folksonomies. A set of 502 users with multiple folksonomy accounts was collected by simply matching the account and person names of Flickr and del.icio.us accounts. This data set was used in the tag-cloud similarity analysis in (Szomszor et al., 2008b) and was also used for populating an ontology for interest and using it for making user recommendations (Cantador et al., 2008). Some of these results are covered in D3.3 and D3.5.

Shortly after we completed the work above, Google Social API was released. We used Google Social API to search each of the 667K del.icio.us users collected last year for their Flickr and Last.fm accounts, if there are any. This search produced 1998 users (users with del.icio.us, Flickr, and Last.fm accounts). We then collected all the tags and tagged resources for each of these users from each of their three accounts. Next we filtered out the users with less than 50 tags in either of their tag clouds, to ensure that the remaining users have sufficient tagging activity for our analysis. The remaining set of users was 1392. This set was used in the work presented in (Szomszor et al., 2008a).

Following the construction of a profile correlation list containing our coplete set of individuals with accounts in del.icio.us, Flickr, and Last.fm, a large set of user data was harvested and converted to an RDF representation. The data collected from each source is summarised below:

- Del.icio.us:
  - Personal information (nickname, homepage url)
  - Posts (including tags and date of post)
  - Friends list
- Flickr:

  - Personal information (nickname, location)
  - Public photos (including geotags, tags, titles, comments, and date)
  - Public contacts
  - Public groups
  - Favourites list

- Last.fm:
  - Personal Information (nickname, location, pictures)
  - List of recent tracks (including artist, album and track)
  - Groups
  - Neighbours
  - Friends

For del.icio.us accounts, complete data for each friend was also harvested, giving a grand total of 6,861 del.icio.us users. For Flickr, any photos in a user's favourites list were harvested, as well as profile information for the user who posted the item. In total, around 8000 individuals were harvested from Flickr. In the case of Last.fm, each neighbour and friend found was also crawled, resulting in a total of over 105 thousand users. When converted to RDF, the complete dataset contains nearly 178 million triples.

Wikipedia category taxonomy was used in the work above for grounding tags to relevant URIs. We are currently investigating using this taxonomy for disambiguating tags. The RDF Wikipedia category, along with other related data (e.g. WordNet synsets, articles) was downloaded from DBPedia and hosted on a triplestore server. The Wikipedia repository amounted to over 50 million RDF triples.

In May 2008 we managed to get a a license to use the JXT[1] triplestore for our research. JXT is arguably one of the most scalable triplestores that currently exists, with a capability of efficiently storing and querying more than 60 billion triples (more if using a cluster of machines). Some of the RDF we have generated in this project is now being transferred to JXT.

In addition to the above, we have built a service for collecting the number of news articles from Google News when given a query term. The purpose of this service is to support our work on investigating correlations between folksonomies. This services allows us to find out how many times the query term was discussed in the news over time. By comparing this trend with tag use in del.icio.us and/or Flickr we might be able to identify some correlations for certain type of tags.

3.1.3 Deviations and Corrective Actions

PHYS-SAPIENZA: none

SONY-CSL: none

UNI KO-LD: The data collection from Tagster is a little behind the planned schedule as the software is not yet as widely distributed as originally planned. The main reason is a higher implementation effort as previously anticipated (cf. 3.2.2). We expect to make up for the delay in the beginning of the third year. Besides, we have some additional data available that was collected from the MyTag platform (cf. D3.5).

UNIK: none

UNI-SOTON: none

[1] `http://esw.w3.org/topic/LargeTripleStores`

3.1.4 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
1.2	(Task 1.2) Data delivery from bibliographic reference sharing systems (Month 23).	1	31 May 2008	31 May 2008	3	3	**UNI KO-LD, UNIK**
1.2	(Task 1.3) Data delivery from experimental tag-based navigation systems (Month 23).	1	31 May 2008	31 May 2008			**SONY-CSL**

3.2 Workpackage 2 (WP2) - Applications

3.2.1 Objectives

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

Task 2.1 Social tagging for online scientific communities:

Task 2.1.1 - Folksonomy web site for sharing of bibliographic data

The work around BibSonomy aimed at further extending the system's functionalities, and at extending the dissemination, in order to attract more users, and thus to enlarge the data on which experiments can be performed.

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

The goal for Tagster during the second year was to extend the application so far that it can be disseminated to a greater community and be used to gather the tagging metadata from representative user group that can then be compared with the datasets collected from delicious and flickr. This included for example improving the navigation support for a better data browsing experience of the user because certain feature like tag clouds were not implemented yet. Additionally, the distributed storing of meta data needed to be improved, too. In the first prototype only a basic distributed index structure was implemented. A more sophisticated mechanism for distributed data management was required for being able to gather certain statistics that can be used for tag cloud generation.

Task 2.2 Tag-based navigation systems

The zexe.net system, in particular the *canal*MOTOBOY* and *GENEVE*accessible* projects, have been included as part of the TAGora applications. The system allows different communities to represent and communicate a commons through mobile, distributed tools. The participants in these projects can send images, videos and sound recordings directly from a multimedia mobile phone to a web page.

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.

3.2.2 Progress

In this section the task responsibles must describe the progresses achieved during the second 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

Within the last project year, we have implemented several features for enhancing the usability of BibSonomy. Feedback on our mailing lists show that these features were relevant for several power users to switch to BibSonomy. The features include: an enhanced group management, extended search functionality (e. g., by author), a concept hierarchy on the personal tag collections, the implementation of the FolkRank ranking, integration support for 3rd party systems (Zope, XWiki, WordPress, Moodle, JabRef), a REST-based web API, encoding support for uploaded files, and multi-language support (English and German).

New features are announced on a weekly basis on `http://bibsonomy.blogspot.com/`.

An important dissemination activity is the Discovery Challenge of the ECML/PKDD conference 2008. This challenge is described in more detail in Deliverable 4.3.

Several university libraries (Köln, Heidelberg, Saarbrücken; Kassel is coming up) have implemented links to BibSonomy into their literature research interface. When searching for books and articles in the library catalogue, the results can be stored with one mouse click in a personal bibliography collection at BibSonomy.

Further dissemination includes a system specific mailing list, cooperation with the Fraunhofer Institute for Autonomous Intelligent Systems and SAP Research, conference support for the Statphys23, ISWC+ASWC 2007, and ESWC 2008 conferences, references from several library catalogues, and publications about the system. Details are described in Deliverable 2.3.

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

In this task we were concerned with improving Tagster's user navigation, implementing distributed statistics to be used for displaying tag clouds, simplifying the network setup, and making the application ready for public dissemination.

User Navigation To improve the user navigation some new view components were implemented. For example, we have integrated the display of tag clouds. Although it is currently only used for showing a user's most frequently used tags it can applied for visualizing any kind of weighted data item set. Moreover, there is a category-based display of resource tags. That means, based on resource's mime-type, we assign the resource's associated tags to the respective categories, image, video, text, music and other. Thus the user has some guidance in his tag-based browsing by narrowing the displayed result list to a specific resource category. Finally, for the resource display we have integrated a hierarchical organized view that directly reflect the tree structure of the local folders the resources are stored in.

Distributed Statistics The basic mechanism for maintaining the distributed tagging meta data is a global index structure that evenly distributes the responsibility for managing data items across all peers in the network. However, this is not sufficient for more complex statistics as frequent communication between many peers would be required and thus the overall message overhead in the system would also drastically increase. Therefore, we have developed a novel mechanism for managing distributed statistics, called PINTS (see (Görlitz et al., 2008)). It is based on the assumption that not all data updates affect all depending statistics which means that only a portion of all update messages needs to be propagated. The challenge, however, is to estimate the effect of an update before it is send. The reason is that the update propagating peer and the peer maintaining the statistic do not know about each other's current data. So if some updates are propagated to late the respective statistics become inaccurate. We have evaluated our algorithm with a simulation using real tagging data taken from the delicious and flickr datasets which were replayed in a peer-to-peer network.

Networking Tagster's networking infrastructure is a very important part of the system. But it is also very complex and requires most of the configuration effort. The user, however, should not really be bothered with bootstrapping and joining of the network. Therefore, Tagster has implemented some automatism to hide most of it from the user. There is, for example, a central user registry that keeps track of all active users in the network. When a new peer joins the network it can contact the registry to retrieve a list of other active peers to connect to.

The use of firewalls in the peer network is still a problem for Tagster. So far we have integrated and tested two different distributed index implementations, Bamboo (`http://bamboo-dht.org/`) and P-Grid (`http://www.p-grid.org/`), which both have their specific advantages and disadvantages, but do not generally support firewall by-passing. Additionally, poor library documentation hampers the effort to integrate such mechanism ourself.

Dissemination and Data Collection The dissemination of Tagster was planned to have two phases. First, the system should be tested within a smaller user group, i.e. the ISWeb research group in Koblenz, to get some feedback about the usability and potential remaining limitations and bugs. In the second step, it should be made availabe for all partners in the project and also to the general public. But due to the encountered problems with the network layer implementation we had to postpone the first phase until the end of the second year. Therefore, also the data gathering started rather late and currently only a relatively small dataset can be made available.

Task 2.2 Tag-based navigation systems

Zexe.net: In the *canal*MOTOBOY* and *GENEVE*accessible* projects, tagging plays a crucial role. Multimedia information can be tagged by its sender, opening thus the possibility for the emergence of folksonomies. *canal*MOTOBOY* started on April 2007, and involves a group of 16 motorcycle messengers in Sao Paulo, Brazil. For this project, the previously existing zexe.net system was completely re-designed and re-written in order to include tagging. We added the possibility of automatically detecting singular and plural forms of tags (based on the rules for portuguese), and grouping these forms into a single bundle. Tag clouds can be viewed according to different criteria: popularity (the size of the tags is proportional to the number of participants who use them) and frequency. Also, when using the tag cloud for performing searches, only the co-occurring tags are highlighted for subsequent filtering.

*GENEVE*accessible* started on February 2008, with the concrete goal of enabling a group of handicapped people to create an online map of the architectural barriers they find in Geneva, Switzerland. For this purpose, the usage of GPS phones and a GIS application were incorporated in the system. In *GENEVE*accessible*, the goal is to correlate tagging data with geographical data in order to generate a dynamic, user-generated cartographic interface.

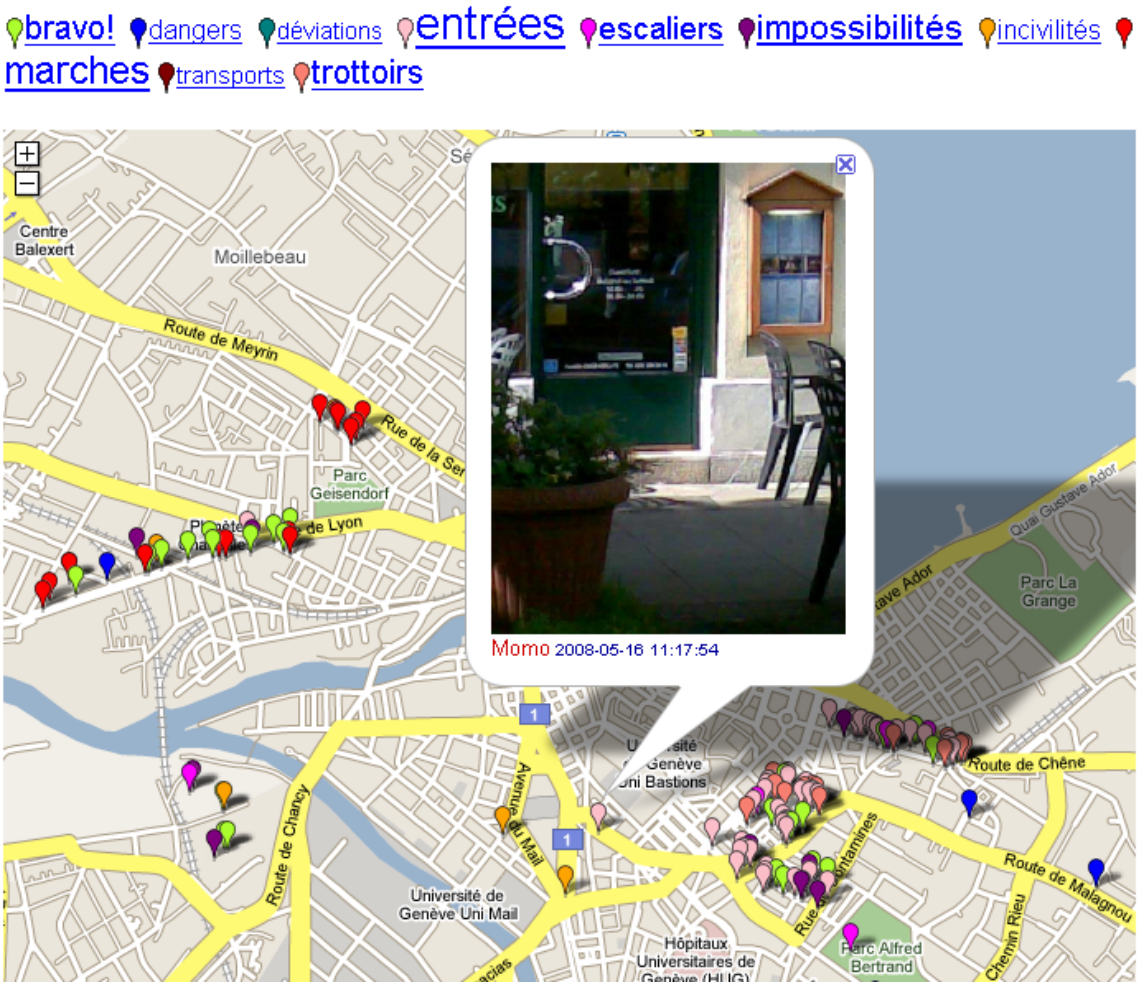

Figure 3.1: GENEVE*accessible

Ikoru: The Ikoru system has been online at `www.ikoru.net` since July 2007. In the past year we have extended the similarity search to music and integrated it in the demo version of the Ikoru web site `demo.ikoru.net`.
We're using Ikoru as a tagging-based backend system for several project. One such project is the use of Ikoru in the artistic installation Phenotypes/Limited Forms. Although the installation is designed to work with the archive of photographer Armin Linke, it is more generally a reflection on museum praxis and how to include visitors' perspectives in the interpretation of the works on display. Another project, which recently started, is to tag musical melodies ((Pachet, 2008)). These differ from audio files in that they are much more abstract, and therefore open to subjective interpretation, but also that they are easier to analyse and generate.

Progress (Milestones)

M2.2 Definition of the control strategy and decision about improvements for the final version of the system for images (Month 23).

Despite it's availability, Sony CSL has decided not to promote the Ikoru Web site to a large audience. It would require too many resources to maintain an active multimedia Web site. In the last two years, many high-quality Web sites integrated tagging for photos. To provide an attractive Web site when popular Web sites can improve their offering at a sustained pace is difficult. However, we will keep the Web site available online and we will target Ikoru for smaller, well-defined projects in which we try to find novel uses for tagging. The improvements that are made to Ikoru through these small projects will be integrated into the main Web site.

M2.3 Definition of the control strategy and decision about improvements for the final version of the system for music (Month 23).

The Ikoru system is used to tag both music and images. We have reached the same conclusions for the system for music as for the system for images in the previous section (M2.2).

3.2.3 Deviations and Corrective Actions

PHYS-SAPIENZA: none
SONY-CSL: none
UNIK: none
UNI KO-LD: The problems experienced with the network layer of Tagster, especially concerning the tunneling of firewalls, has delayed the planned dissemination to a larger community. We expect to implement a workable solution by the beginning of the third year to quickly catch up with the original schedule for gathering the tagging data of a larger user group.
UNI-SOTON: none

3.2.4 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
2.3	(Task 2.1, 2.2., 2.3) Interim report on tagging systems update and usage (Month 23).	2	15 Jul 2008	20 Jun 2008	2	2	**UNIK (ALL)**

Mil. No.	Milestone name	WP No.	Date due	Actual/ Forecast delivery date	Lead contractor
M2.2	Definition of the control strategy and decision about improvements for the final version of the system for images (Month 23).	2	15 Jul 2008	20 Jun 2008	**SONY-CSL**
M2.3	Definition of the control strategy and decision about improvements for the final version of the system for music (Month 23).	2	15 Jul 2008	20 Jun 2008	**SONY-CSL**

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

3.3.1 Objectives

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

During the first year of the project, the main focus of WP3 was on gathering emergent metadata statistics and the identification of relevant topological and dynamical properties of tagging systems. During the second year, these two activities were continued but the main focus shifted towards the other WP3 tasks, i. e. the identification of clusters/communities (T3.3), the incorporation of background knowledge for using semantic inference (T3.4) and the identification and analysis of cross-folksonomy networks (T3.5). These tasks make use of the already collected statistics and properties and will help to further improve the theoretical analysis and understanding of folksonomies but they can also be used for providing practical tools for improving tagging systems.

- In T3.3 we started with identifying community structures within tagging systems. The identification of communities is an important step for getting a better understanding of the influence of social relationships on the complex dynamics in tagging systems but it may also improve the results of recommender systems (cf. WP4).

- In T3.4 we incorporated sources of background knowledge into the analysis of tagging data in order to improve the retrieval of resources from the systems. The integration of background knowledge may e. g. help in overcoming one of the major problems during resource retrieval from tagging systems, namely the sparsity of tagging data. But it can also be used

for cleaning the raw tagging data, e. g. by merging synonyms or different abbreviations of the same concept.

- In T3.5 we started to identify and analyze cross folksonomy networks. This may e. g. help in understanding how different tagging systems influence each other. But besides this theoretical interest in cross-folksonomy networks it will also have a practical impact on applications like cross-folksonomy recommendations.

Task 3.1 Emergent metadata statistics

Task 3.1 deals with the quantitative statistics coming from the analysis of the datasets provided by WP1. As a first approach, this includes basic information like the number of users, tags resources and tag assignments but also frequency distributions of tags and number of tag assignments per user.

While in the first year we mainly analyzed all those static statistical properties, during this year we focused our investigation on dynamical statistical properties in order to better characterize the co-operative behavior of users, if any. We investigated the dictionary growth, the tag-tag correlations and the inter-arrival times of tags inside tag streams.

Further, we analyzed simultaneously the temporal evolution of tag streams extracted from different folksonomies, looking for sinchronous user activity in correspondence of known external events. Also in this spirit, the occurrence of a set of tags with similar meaning (eg. “H5N1” and “avian flu”) over time were monitored in order to unravel correlated users activity.

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 that are available 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 identification of communities (see T3.3). Already in the first year, specialized measures for the tri-partite structure of folksonomy networks were developed.

In the second year the main objective is to further analyze specific projections in order to reveal correlations (semantic or social) emerging from collective tagging activity. The emerging features, should inspire the corresponding modeling activity.

Task 3.3 Cluster/community identification

This task summarizes our work on community detection in folksonomies. Our work can be grouped into three parts. We present an approach that starts with the generation of a hierarchical clustering of the tag space by iteratively applying the k-Means clustering algorithm. The tag clusters on the bottom level are then considered as intensional descriptions of our FolkRank algorithm. For the choice of k-Means as initial clustering algorithm, we provide the semantic grounding, which shows that the average semantic distance of pairs of tags within clusters generated by k-Means is significantly smaller than within randomly generated clusters.

A similar approach has been implemented in BibSonomy. Here, the focus lay on efficiency, since we display the results online on all tag pages. We provide, for each tag, the community of users that are mostly related to this tag.

Task 3.4 Semantic Inference

The analysis of the collected datasets can greatly benefit from the use of background knowledge because relations between data items may only become apparent on the semantic level. For ex-

ample, synonyms or tags in other languages can only be partially identified by an analysis of the tag networks in folksonomies. Furthermore, background knowledge may bring in taxonomic information. Sources of background knowledge may for example be basic dictionaries but also more complex ontologies. In this task we investigate the usefulness of different sources of background knowledge for processing the collected datasets. Additionally, it is studied in how far it may be used for doing semantic inferences and for applying machine learning on tagging data.

Task 3.5 Cross-Folksonomy Networks

This task aims to bridge between some of the data we collected from multiple folksonomies to facilitate researching (a) tag evolution across folksonomies, and (b) cross-folksonomy recommendations. To reach these goals, we need to first understand how tags from different folksonomies or from individual tag clouds can be mapped to each other. This involves investigating how to filter those tags, disambiguate them, and how to ground them to URIs to build the integrated semantic network that covers multiple folksonomy data. Once the data is integrated, a network of tags, users, and resources will be created which can be explored for investigating the issues of tag evolution and cross domain recommendations.

Task 3.6 Collaborative tagging and emergent semantics

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.

3.3.2 Progress

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

Task 3.1 Emergent metadata statistics

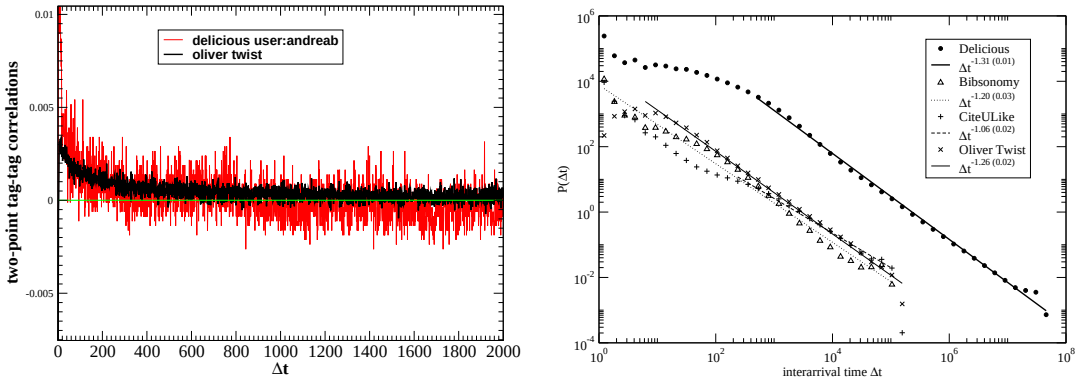

Figure 3.2: *Left:* Tag-Tag correlation for the del.icio.us user *AndreaB* compared with the word-word correlation of Dickens' novel *Oliver Twist*. Curves where shifted in order to share the same asymptotic poissonian value. *Right:* Tag inter-arrival time distribution in collaborative tagging communities compared with word inter-arrival times in Dickens' novel *Oliver Twist*.

To complete the stream analysis performed in the first year, an extensive description of the dif-

ferent properties of co-occurrence and resource streams was created, with particular attention to frequency distributions and growth of tag dictionary size. Subsequently, the emergence of these properties were successfully explained with the help of the suggested epistemic dynamic model for tagging systems.

Moreover, the stream analysis hasn been concentrated on the measure of time correlations. This has been achieved considering two-times correlations, as well as distribution of inter-arrival times of tags. The measures reveales non trivial correlations and has been compared to same measures performed in book texts, as well with random shuffled streams. In Figure 3.2 is shown example measures, respectively, of two-times correlation and inter-arrival time distributions (Capocci et al., 2008).

Task 3.2 Network/graph analysis:

With regard to analyzing the topological properties of networks, Koblenz submitted an overview article about social networks for publication in the Database Encyclopedia. It contains an overview of different notations for social networks as well as related measures for analyzing e. g. their socio-centric or ego-centric properties and for identifying subgroups within the networks. Furthermore, it lists key applications of social network analysis, possible future directions and available data sets that are widely used in the literature (e. g. the Enron e-mail or the DBLP dataset).

Current measures for social network analysis can only be applied on networks which only contain one kind of relationship between the different actors. Thus, in the context of (Henkes, 2008) it was explored in how far one can modify the existing measures so that they can distinguish the different relationships available in real social networks. For example, a social network may not only contain a general *knows* relationship but it may be further differentiated into e. g. *isFriendOf* or *isColleagueOf*.

For experimenting with the the modified measures we simulated social networks where each actor has assigned certain topics on which he/she is expert. In this network, it was then differentiated between several *talks about topic X with* relationships. The simulation was based on the REMINDIN routing algorithm described in (Tempich, 2006) and that was designed for routing messages in semantic peer-to-peer systems. The simulation of the relationships was initialized with the real distribution of topics to authors that is available in the DMOZ dataset of the Open Directory Project[2].

During the experiments, the modified measures were then used for studying how the properties of social networks change if one restricts the analysis to specific relationships in the graph. The experiments also contained a dynamic component, i. e. during the simulation of the social networks the relationships were initialized by the small-world graph generator described in (Kleinberg, 2000). It was then analyzed in how far different simulation parameters change the initial relationships between the actors as well as how the topic specific relationships between actors are affected or whether different properties can be observed, compared to a network with only general relationships.

The study of the tag co-occurrence network revealed a rich structure, sign of semantic correlations between tags. For instance we considered the co-occurrence built from the tag co-occurrence stream, which should limit the semantic context, and measured several statistical quantities (see Figure 3.3). A model, described in T4.1, succesfully reproduces these measures, as well the growth of the tag dictionnary size in the stream.

Many other quantities have been measured, revealing non trivial correlations in the tag co-occurrence network. For instance, the k-core structure of the network, as well as the analogous measure taking account of the co-occurrence weights, are shown in Fig. 3.4 compared with a shuffled co-occurrence network, where the semantic correlations has been destroyed. As can be seen in the Figure, the non shuffled network shows a reduced size of the cores, indicating a different

[2] http://www.dmoz.org/

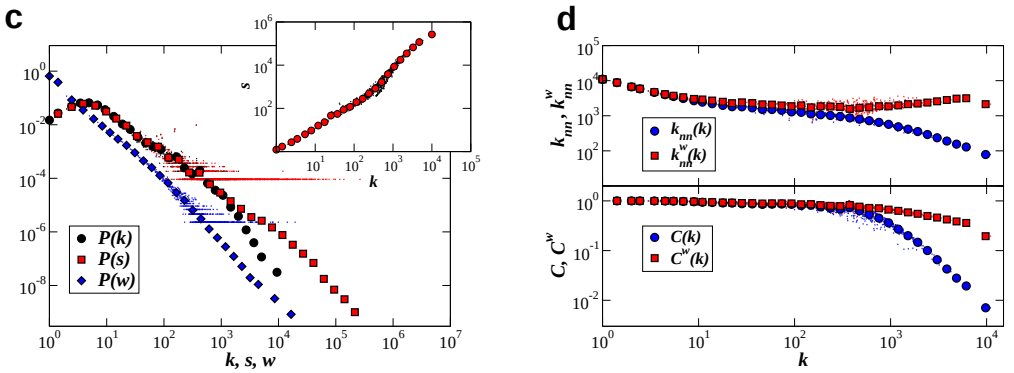

Figure 3.3: *Left:* Broad distributions of degrees k, strengths s and weights w are observed. The inset shows the average strength of nodes of degree k, with a superlinear growth at large k. Both raw and logarithmically binned data are shown. *Right:* Weighted (k^w_{nn}) and unweighted (k_{nn}) average degree of nearest neighbors (top), and weighted (C^w) and unweighted (C) average clustering coefficients of nodes of degree k. k_{nn} displays a disassortative trend, and a strong clustering is observed. At small k, the weights are close to 1 ($s(k) \sim k$, see inset of B), and $k^w_{nn} \sim k_{nn}$, $C^w \sim C$. At large k instead, $k^w_{nn} > k_{nn}$ and $C^w > C$, showing that large weights are preferentially connecting nodes with large degree: large degree nodes are joined by links of large weight, i.e. they co-occur frequently together. Both raw and logarithmically binned data are shown.

network topology and suggesting a hierarchical organization of tags.

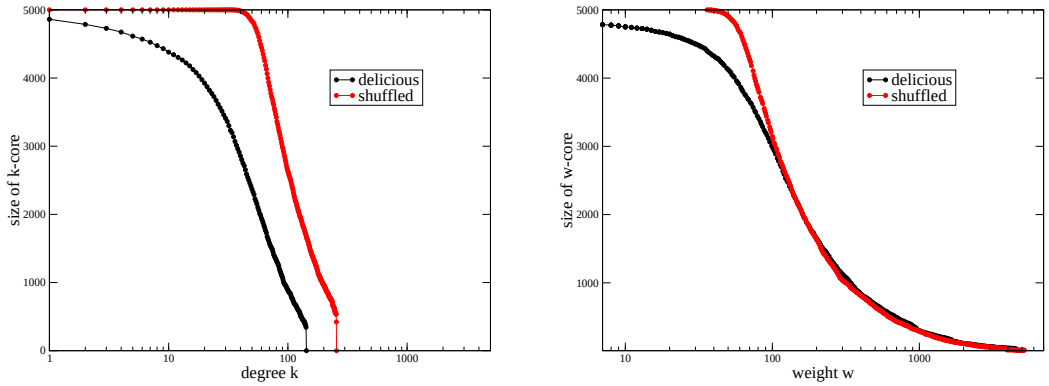

Figure 3.4: *Left:* k-core analysis of the co-occurrence network obtained with the first million of posts in the delicious tag stream; the red curves represents the network obtained with a shuffled stream, where the post structure has been kept, while the tags have been randomly permuted. *Right:* core analysis performed taking in account the weighted nature of the network. Again the measure is repeated on a shuffled stream, where semantics correlations are absent.

Task 3.3 Cluster/community identification

In Deliverable 3.2, we discuss three different studies on community detection which we performed within the project:

The first approach starts with the generation of a hierarchical clustering of the tag space by iteratively applying the k-Means clustering algorithm. The tag clusters on the bottom level are then considered as intensional descriptions of our FolkRank algorithm. For the choice of k-Means as initial clustering algorithm, we provide the semantic grounding, which shows that the average semantic distance of pairs of tags within clusters generated by k-Means is significantly smaller than

within randomly generated clusters.

A similar approach has been implemented in BibSonomy. Here, the focus lay on efficiency, since we display the results online on all tag pages. We provide, for each tag, the community of users that are mostly related to this tag.

Task 3.4 Semantic Inference

For semantic inferencing based on background knowledge it was first necessary to investigate what background knowledge would be useful for which tasks. We focused on different aspects for using background knowledge, namely for data filtering, data enrichment and data classification.

- Data filtering: Pre-processing the tagging dataset before analysis, i.e. splitting compound words, merging singular and plural forms, stemming etc., is often very useful for obtaining better results. Different sources of background knowledge were investigated to support that task, e.g. Wordnet, Wikipedia, and Google search, and integrated in a filering architecture.
- Classification: In D3.1, T-Org (see (Abbasi et al., 2007)) was presented as one approach to classify tags into predefined categories. In the second year, other classification approaches were investigated. For example, Flickr tags contain a high number of location related annotation. We have developed the Triple Play approach (see (Abbasi et al., 2008)) using the Geonames database for identifying location names used as tags and a SVM to classify the tags. Additionally, classification of audio data was done by combining audio classifiers with results obtained from tag correlations.
- Semantic Analysis: Another problem for inferencing knowledge from the datasets is the sparseness of data that influences search results. Methods like Triple Play (see (Abbasi and Staab, 2008)) were developed for overcoming this problem by recombining the existing data and using them in Latent Semantic Analysis to identify hidden concepts.

Task 3.5 Cross-Folksonomy Networks

During the second year of TAGora, work on Task 3.5 was intensified, where we started investigating cross-linking tag clouds. Work concentrated on four main efforts:

- Tag filtering: After scrutinizing a sample of tag clouds, it was clear that tags needed to be filtered to increase tag-cloud compatibility. In other words, tags had to be filtered to consolidate synonyms, remove meaningless tags, break up compound tags, etc. Filtering tags improves their comparisons and facilitates grounding them into appropriate URIs. To this end, we implemented a tag filtering process that receives a raw tag cloud and produced a filtered set of tags. The process is fully explained in (Cantador et al., 2008) and was used in the next two items of work. This process is also summarized in D3.3.
- Tag cloud similarity: To better understand how users tag across folksonomies, we carried out an experiment with 502 users who happened to have an account in del.icio.us and an account in Flickr, to study the similarity between their separate tag clouds. The results showed that the tag cloud of a user from del.icio.us tend to be more similar to the Flickr tag cloud of this same user than to other Flickr tag clouds. This experiment showed that users tend to carry their tagging trends and choices across folksonomies, even when the folksonomy is for a completely different domain. These results are fully covered in (Szomszor et al., 2008b).
- Mapping tags to Wikipedia categories: To use cross-folksonomy integration for recommendation purposes, we investigated using Wikipedia category to ground tags into URIs to represent their true concepts. These concepts are then regarded as the interests of the tagger,

and used to generate a semantic profiles of interests. These profiles are built using FOAF and Wikipedia category taxonomy. Results so far indicate that much more can be learnt about users when expanding the analysis of their tagging activities to other folksonomies. These results are explained in D3.5 and can be found in (Szomszor et al., 2008a).

- Correlation between folksonomies: We started analyzing del.icio.us, Flickr, and Google News to find out if they correlate in terms of tag usage trends. We are still at the stage of experimentation with various analyses algorithms to find the most suitable one to use for this purpose. This work will continue into year 3 of the project.

We are currently investigating an approach for disambiguating tags based on the position of corresponding concepts in Wikipedia category hierarchy. Tag disambiguation is a necessary step towards better understanding and representing user interests.

In addtion to that a new web application, a cross folksonomy search and recommendation tool, was set up by the University of Koblenz team `http://mytag.uni-koblenz.de/`. MyTag allows users to search accross folksonomy sites YouTube, Flickr and del.icio.us. Additionally, users can create a personal profile on MyTag that is used to tailor the search results to the users personal interests. This application was developed within the context of TAGora, and is being used to disseminate the project through the software implementation of theoretical research. MyTag is being used to track user interests and to gather search requests that can later be used for further analysis.

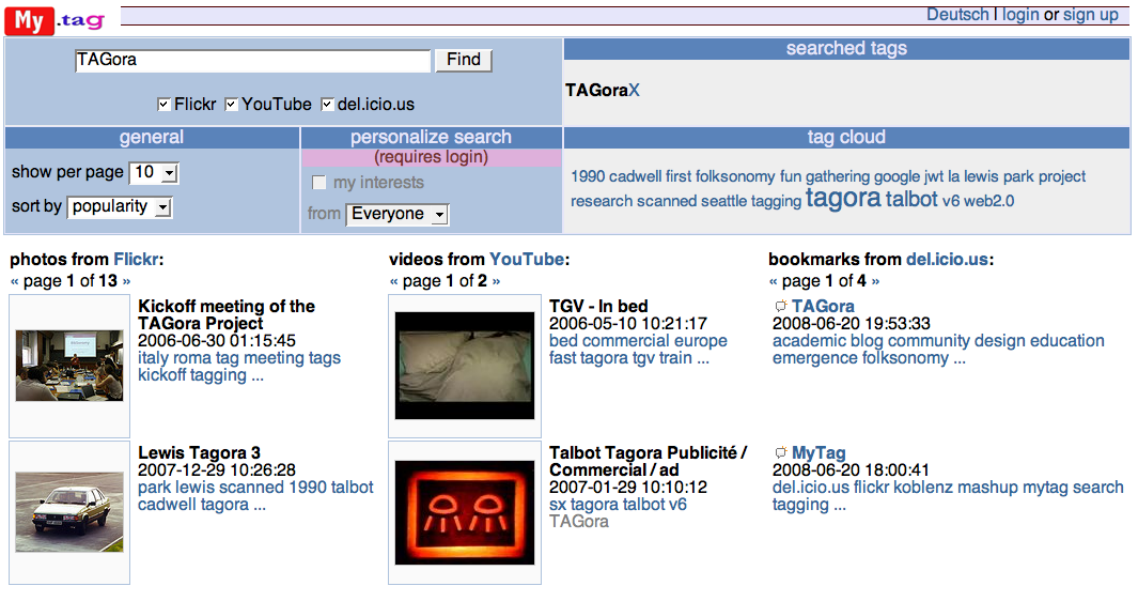

Figure 3.5: Screenshot of MyTag

Task 3.6 Collaborative tagging and emergent semantics

Sony CSL conducted intensive and rigorous experiments regarding the automatic classification of acoustic signals with respect to tags describing a music title in its entirety, such as its genre, mood, main instruments or type of vocals. This is a *supervised-learning* task that is typically addressed by training a classifier for each tag. The classifiers are trained on feature values that are computed for each title, and they learn the tags that set by humans (the so-called *ground-truth*). The performance of such individual classifiers (i.e. modelling a *single* tag) are rarely satisfactory.

For music collections in which the titles have multiple tags, we have introduced the *correction hypothesis*, which postulates that it is possible to exploit existing redundancies between tags to correct some of the errors of individual acoustic classifiers.

We introduced an implementation of this hypothesis, the *correction approach*, whereby, for each tags, a *correction* is trained on the output of all the individual acoustic classifiers. We conducted a series of experiments to validate this hypothesis on a large-scale database of music and metadata (32,000 titles and 600 boolean attributes per title). The experiments validate the hypothesis and highlight several interesting phenomena such as the feature independence of the correction.

We found that the *correction hypothesis is true*. On average, correction classifiers perform better than the corresponding acoustic classifiers and the correction approach provides an almost-systematic performance improvements. In addition, the improvements are feature-set independent: when we use two distinct feature sets, we observe a strong parallelism between the performance improvements of both. More details on this study can be found in Deliverable 3.3 and in (Pachet and Roy, 2008a,b,c; Rabbat and Pachet, 2008).

3.3.3 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-KOLD: none

UNIK: At UniK, a central researcher (M. Grahl) resigned from her contract end of March 2008. Because of the recently started national initiative for excellency which created many open positions, no adequate replacement could be found up to now. The staff situation reduced the effort we could invest, especially for Deliverable 3.2. To partially compensate, we shifted our focus on the topics of logsonomy analysis and spam detection – where internal expertise was available.

UNI-SOTON: Studying how tags evolve and spread across folksonomies was one of the aims of Task 3.5. After some analysis and initial experimentation, it became clear that such an integration process should not be performed without first implementing the tools and procedures for cleaning the tags to ensure that tags are correctly mapped to each other. Failing to do so can produce incorrect integrations which may damage our results. To this end, work during the past year was focussed on implementing various tag filtering and matching techniques as described earlier in this report, and detailed further in D3.5. Having implemented these tools, and gathered large amounts of data as described in section 3.3.2 Task 3.5, we are now in a good position to investigate cross-folksonomy tag evolution. We have already studied and tested a number of algorithms for detecting correlations, and found that the most promising algorithm to use is the one described in (Kleinberg, 2002).

3.3.4 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
3.2	Methods for identifying communities (Month 23).	3	15 Jul 2008	20 Jun 2008	1	1	**UNIK (ALL)**
3.3	Methods for using semantic inference in data analysis (Month 23).	3	15 Jul 2008	20 Jun 2008	2	2	**UNI KO-LD**
3.5	(Task 3.5) Protocol for integrating cross-folksonomy networks (Month 23).	3	15 Jul 2008	20 Jun 2008	2	2	**UNI-SOTON**

3.4 Workpackage 4 (WP4) - Modeling and simulations

3.4.1 Objectives

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

Task 4.1 Modeling: Social bookmarking systems allow users to store links to internet resources. The structure behind these social systems, called folksonomies, can be viewed as a tripartite hypergraph of user, tag, and resource nodes. The graph shows specific topological properties that explain its growth and the possibility of serendipitous exploration; characteristics that have contributed to the recent raise in popularity of folksonomies. The modeling activity of the project will proceed in refining the results obtained in the first year (modeling of average user behavior and descritption of tag streams) but it will be oriented through more ambitious objectives.

Folksonomy structure. More specifically on the folksonomy systems, the WP4 is devoted to the analysis and modeling of the very structure of folksonomy. We would like to better reveal and understand correlations derived from users activities: these can be originated from the common background knowledge (for instance coming from tag semantics) or from temporal patterns in tagging activity. The study will be directed mainly in these two directions: analysis and modeling of the tag co-occurrence network, as well as the study of temporal tag streams.

Folksonomy usage. As more and more people add content, social bookmarking systems present an alternative to traditional web search engines. Considering this new information source, several questions arise: Do people use search engines and social bookmarking systems in a similar way? Is the content different? Can the feedback of web search engine users provide similar information as the explicit tagging of a web resource?

Task 4.2 Control :

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

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

Task 4.2.2 - Ontology learning

Ontology learning tries to capture the concepts inherent to tag annotations in the folksnonomies. However, they are also significantly depending on the social relations among users and how they

influence each other. To improve the identification and learning of concepts it is necessary to take these social aspects into account. Therefore, we need to understand how and when users influence each other, and what properties are significant for propagating concepts in the social network.

Task 4.2.3 - Simulation and control on bibliographic reference sharing system

Usual counter-measures like captchas are not efficient enough to effectively prevent the spamming of systems like BibSonomy. One possibility to approach this problem is to classify users as spammers resp. non-spammers by machine learning algorithms.

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.

3.4.2 Progress

Task 4.1 Modeling:

Folksonomy structure. The model introduced in (Cattuto, 2006; Cattuto et al., 2006, 2007b) opened the way to modeling tagging user behavior, in order to explain the main statistical features of tag streams, notably the tag frequency distribution. In the second year of activity, Koblenz proposed an epistemic dynamic model (a more refined version of the stream model proposed in (Cattuto, 2006)) that can be used for simulating the assignment of tags to resources (see Task 4.1 and (Dellschaft and Staab, 2008)). It assumes two different main influences on the user during assigning tags: (1) The imitation of previous tag assignments made by other users and (2) the selection of tags from his background knowledge that he thinks are suitable for describing the content of the resource. We successfully manage to describe and understand the observed differences between different kind of tag streams and to reproduce the characteristic sub-linear growth of the tag dictionary size (Cattuto et al., 2007a). Furthermore, the model also identify the parameters that influence the dynamics, namely the probability with which a user imitates previous tag assignments and the probability with which a tag comes from the background knowledge of the user. A software simulator has been developed to describe the model and to produce synthetic tag streams (see Fig. 3.6). Further analysis of tag correlation has been carried on, in particular comparing tag

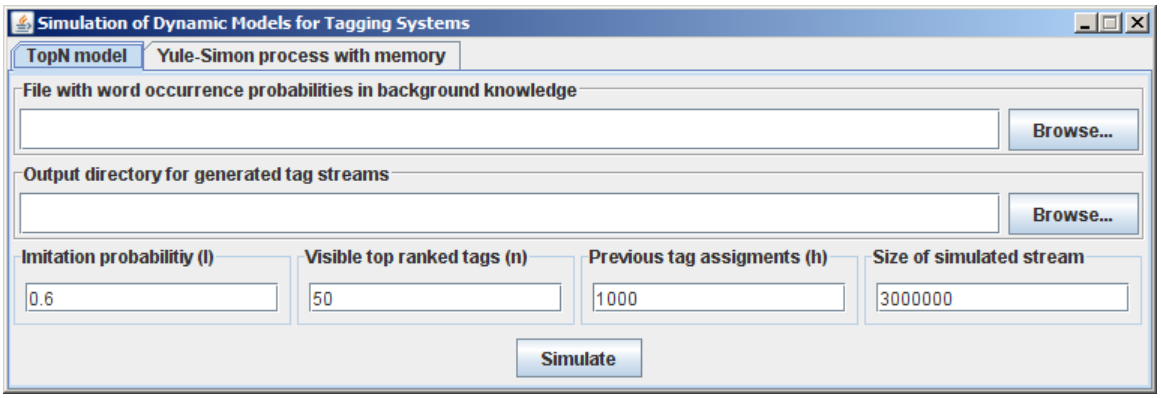

Figure 3.6: Screenshot of the simulator interface.

streams and book texts. Furthermore an analysis of inter-arrival time of tags allowed a scaling analysis (standard in statistical physics, but absolutely not obvious in the context of human activity dynamics) which reveal again non trivial correlation in user tagging activity, i.e. the presence of temporal pattern. Even though these investigations show the relevance of tag stream analysis, the full richness of the folksonomy is contained in its network structure (Cattuto et al., 2007c). A specific projection, the tag co-occurrence network, encodes semantic relations between tags, and its study opened the way to further models and new control strategies in applications. We investigated several notions of node (resource or tag) similarity in a folksonomy, and characterized them semantically (Cattuto et al., 2008c; Markines et al., 2008). Our results pointed out that a specific similarity measure (Cattuto et al., 2008b) (cosine similarity between tag co-occurrence vectors) can spot "synonym" or sibling tags. An efficient technique to compute this measure was devised and implemented in BibSonomy, enhancing its navigation interface to include recommendation of "similar" tags. As it turns out (Cattuto et al., 2008a), the distribution of pair similarities between tags is a able to reveal non-trivial correlations in the tag co-occurrence network, ascribable to semantics. This result was used to build an accurate model of the network structure of the tag co-occurrence network.

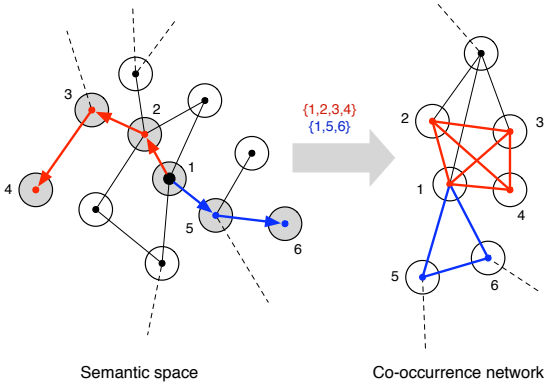

Figure 3.7: Illustration of the proposed mechanism of social annotation. The semantic space is pictured as a network in which nodes represent tags and a link corresponds to the possibility of a semantic association between tags. A post is then represented as a random walk on the network. Successive random walks starting from the same node allow the exploration of the network associated with a tag (here pictured as node 1). The artificial co-occurrence network is built by creating a clique between all nodes visited by a random walk.

The model is based on the idea that a shared (latent) semantic space drives the tagging activity of users. Accordingly, a latent graph of tags is used as a space for (short) random walks (see Fig. 3.7), in order to produce synthetic posts. The resulting tag stream reveal, again, a broad frequency distribution and the typical sub-linear growth of the dictionary size is reproduced. Moreover, from the generated posts, a synthetic co-occurrence network is derived. Interestingly, this network shares many of the statistical features observed in real data (included the distribution of similarities between tags), and this result is very robust with respect the topology of the latent semantic graph.

Folksonomy usage. In a study, we have compared search in social bookmarking systems with traditional web search. In the first part, we compared the user activity and behaviour in both kinds of systems, as well as the overlap of the underlying sets of URLs. Our experiments are performed on data of the social bookmarking system Del.icio.us and on rankings and log data from Google, MSN, and AOL. We have shown that part of the difference between the systems is due to different behaviour (e. g., the concatenation of multi-word lexemes to single terms in Del.icio.us), and that real-world events may trigger similar behaviour in both kinds of systems. In the second

part of the study, we have transformed search engine logs into the structure of folksonomies. The structure of the resulting 'logsonomy' can then directly be compared to a folksonomy of a social bookmarking system. In particular, we have compared logsonomies from MSN and AOL search logs with a snapshot of the folksonomy of the bookmarking system Del.icio.us. The results show that folksonomies and logsonomies indeed have similar characteristics. Our findings suggest to re-use social aspects which arise from the resulting network structure of a folksonomy in search engines as well as to use information retrieval techniques to improve search in folksonomies.

This work has been presented at the 30th European Conference on Information Retrieval (Krause et al., 2008b) and at the Second International Conference on Weblogs and Social Media (Jäschke et al., 2008). An extended journal version is under submission.

Task 4.2 Control:

Task 4.2.2 - Ontology learning

The behavior of users and the social influences among them in the folksnonomy is important for ontology learning. Therefore, understanding the basic factors driving the users tagging activities was of special interest during the second year. The tagging models developed in Task 4.1. (cf. D4.2) provide such kind of fundamental insights.

The epistemic dynamic tagging model by (Dellschaft and Staab, 2008) generates a tagging stream that shows the same characteristics as found in the Flickr and delicious datasets. It takes into account different factors, i.e. tag imitation, invention, and the use of background knowledge. Several experiments with different values for these factors were conducted to verify the model and compare it with the characteristics found in the gathered data sets. The results show that the model is able to reproduce the typical characteristics quite closely. We expect this knowledge to be helpful in further improving the ontology learning process.

Task 4.2.3 - Simulation and control on bibliographic reference sharing system

The classical approach in machine learning to a classification task like spam detection is to determine relevant features that describe the system's users, train different classifiers with the selected features and choose the one with the most promising evaluation results. We have transferred this approach to a social bookmarking setting to identify spammers. We have presented features considering the topological, semantic and profile-based information which people make public when using the system. The dataset used is a snapshot of the social bookmarking system BibSonomy and was built over the course of several months when cleaning the system from spam. Based on our features, we have learned a large set of different classification models and compare their performance. Our results represent the groundwork for a first application in BibSonomy and for the building of more elaborate spam detection mechanisms. This work has been presented at the Fourth International Workshop on Adversarial Information Retrieval on the Web 2008 (Krause et al., 2008a).

Spam Detection on BibSonomy is also a dissemination activity of the TAGora project. This year, we organise the Discovery Challenge of the ECML/PKDD conference, see `http://www.kde.cs.uni-kassel.de/ws/rsdc08/`, which addresses spam detection and tag recommendations in BibSonomy.

Task 4.2.4 - Recommendations based on Network Analysis

The most important stage of building a recommender system is gathering and processing data about users. This was the focus of our work in this task in the past year of the project. We have collected information about users with multiple folksonomy accounts (del.icio.us, Flickr, and Last.fm) and mapped their tags to Wikipedia categories to ground their URIs. The result was a semantic profile of each of these users, using the FOAF and Wikipedia category ontologies.

The first experiment was aimed at understanding how similar user tag clouds tend to be across the various folksonomies. To this end, a set of 502 users with del.icio.us and Flickr accounts was collected and the similarity of their tag clouds was measured. Through this investigation, we

discovered that prominent user interests, important locations, and events, are often reflected in the intersection between tag-clouds, irrespective of the focus of the folksonomy. We also found that the multiple tag clouds for a user tend to be more similar to each other (i.e. have more overlap) than to other user's tag clouds. This work is detailed in (Szomszor et al., 2008b).

We then experimented with populating a set of 17 ontologies which were subsets of the IPTC ontology[3] for news classification with information gathered from 502 users. The aim 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. Over 137 thousand Wikipedia entries were used to populate 744 ontology concepts with over 121 thousand instances. Each instance represents a concept that matches a tag used by a user in either del.icio.us or Flickr. Evaluating these results showed that the average accuracy of class assignments was 69.9%, with 84.4% average accuracy of ontology instantiation.

The ontology profiles above were then used to make recommendations of news to the corresponding users. These recommendations where then evaluated and showed an average accuracy of 67%, in comparison to only 39% when relying on keywords only. Full detail of this work can be found at (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. We needed to know how much new knowledge can we expect to gather about users if we expand their profiling to include their Flickr accounts in addition to their del.icio.us accounts. Over 1390 users with accounts in del.icio.us, Flickr, and Last.fm were collected and used for this experiment. The results showed that on average, 15 new interests were learnt for each user when expanding tag analysis to their tag cloud in the other folksonomy. This experiment also highlighted a number of issues that need further investigations, such as tag disambiguation, specificity and frequency of user interests, indirect inference of interests, and tags that map to several Wikipedia categories. All these issues will be thoroughly researched in the third year of the project. Full detail of this experiment and work is described in (Szomszor et al., 2008a).

Progress (Milestones)

M4.1 (Task 4.1) Adoption of a set of models that capture the essence of the emergent behavior and describe them (qualitatively and, whenever possible, quantitatively) (Month 17).

Recent research has brought forward an interesting temporal perspective on the understanding of folksonomies by viewing them as dynamic stochastic systems with memory. But this perspective abstracts away the background knowledge common to folksonomy users putting too much emphasis on imitation of other users and random generation of vocabulary. We advocate the hypothesis that both components, i.e. the background knowledge and the imitation, are needed for explaining and understanding the tagging behavior of users.

At the same time we extended our modeling activities to more complex structures in folksonomies. We focused in particular on a generative model for the co-occurrence network. In particular we have shown that the process of social annotation can be seen as a collective exploration of a *semantic space*, modeled as a graph, through a series of random walks. Strikingly these simple assumptions reproduce several main 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.

[3] http://nets.ii.uam.es/neptuno/iptc/iptc-srs.rdfs

M4.2 (Task 4.2) Implementation of realistic simulation software aimed at the control experiments (Month 17).

A generative tagging simulator has been developed by the Institute for Computer Science at the University of Koblenz-Landau. It integrates both the background knowledge and the influence of previous tag assignments. It successfully reproduces characteristic properties of tag streams and even explains effects of the user interface on the tag stream. The simulator employs two widely accepted algorithms for producing artificial tag stream, statistically similar to the real datasets found in collaborative tagging communities. It is written in Java and is available online via the institute's website, where additional documentation can be found. Along with the software, an archive containing all generated tag streams, the software simulator and the technical report is provided. This simulator can be used to test many different modeling schemes before implementing control experiments.

M4.3 (Task 4.1) Feedback to WP2 about the best control strategies inspired by the modeling and the simulation activities (Month 23).

During the second year we started the process of feedback from data analysis and modeling to applications and tools. Several new features inspired or enabled by our previous research have been implemented in BibSonomy. In particular, we introduced navigation aids stemming from our research on semantically-grounded notions of tag and resource "similarity". Our FolkRank algorithm has been deployed and now it allows to identify related users. Finally, even though a first implementation of a spam detection system already exists in BibSonomy, we are actively implementing the best spam detection strategies, as determined by our experiments.

M4.4 (Task 4.2) Implementation of a semantic recommender (Month 23).
As mentioned earlier, we have been actively collecting and integrating information gleaned from multiple folksonomies to create a large semantic network of users, tags, and resources. We have performed a number of experiments and tests to verify our approaches on tag cloud integration, tag cleaning, and URI assignments. We are now planning the development of a web application for users to access and visualize the results from merging their distributed tag clouds. This application will probably combine the services and tools developed as part of Task 3.5 and Task 4.2.4 as well as others, with MyTag (MyTag is described in D3.5) to allow users to view and evaluate the results, and also to receive recommendations of similar items and/or users. The first release of a prototype for a semantic recommender system based on folksonomy data is planned for October of this year.

3.4.3 Deviations and Corrective Actions

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

PHYS-SAPIENZA: none
SONY-CSL: none
UNI KO-LD: none
UNIK: none
UNI-SOTON: The implementation of a recommender system was slightly delayed due to unexpected research challenges (tag filtering, disambiguation, URI selection) which required further

work and analysis before such a system can be developed. We have secured additional person effort this summer and will be starting on implementing this system in July of this year.

3.4.4 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
4.2	(Task 4.1) Interim report describing the models and the simulation schemes selected and/or developed in order to quantitatively describe the observed emergent properties and the insights gained by comparing models and actual systems (Month 23).	4	15 Jul 2008	20 Jun 2008	2	2	**PHYS-SAPIENZA (ALL)**
4.2	(Task 4.1) Report on the roadmap leading from modeling activity to control strategies (Month 23).	4	15 Jul 2008	20 Jun 2008			**PHYS-SAPIENZA (ALL)**
4.3	(Task 4.1) Set of software simulators implementing the best performing modeling schemes and the ensuing control strategies (Month 23).	4	15 Jul 2008	20 Jun 2008	2	2	**PHYS-SAPIENZA**

Mil. No.	Milestone name	WP No.	Date due	Actual/ Forecast delivery date	Lead contractor
M4.1	(Task 4.1) Adoption of a set of models that capture the essence of the emergent behavior and describe them (quantitatively and, whenever possible, quantitatively) (Month 17).	4	30 Nov. 2007	30 Nov. 2007	**PHYS-SAPIENZA**
M4.2	(Task 4.2) Implementation of realistic simulation software aimed at the control experiments (Month 17).	4	30 Nov. 2007	30 Nov. 2007	**PHYS-SAPIENZA, UNIK**
M4.3	(Task 4.1) Feedback to WP2 about the best control strategies inspired by the modeling and the simulation activities (Month 23).	4	31 May 2008	31 May 2008	**UNIK**
M4.4	(Task 4.2) Implementation of a semantic recommender (Month 23).	4	31 May 2008	30 July 2008	**UNI-SOTON**

3.5 Workpackage 5 (WP5) - Dissemination and exploitation

3.5.1 Objectives

The objectives for the second year is the same as the first year: to disseminate the research results, applications and strategies generated by the TAGora project within scientific and artistic communinties, and also to communicate and apply them to a wide, general audience.

Task 5.2 Dissemination strategies

During the second year of activity, the members of TAGora have focused an important part of their efforts in disseminating the outcomes of their research and the products developed within the project. As a result of this activity, the TAGora project is quickly 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 (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 inside and oustide the Web, includes the following:

- Publication of research done within TAGora scientific journals and other written communication media, aiming to reach both the scientific community and the general public.
- Presentation of research results in different types of scientific and artisitic events, such as conferences, talks, workshops, courses, demos and axhibitions.
- Maintainance 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.

Task 5.2.2 - The role of applications developed in WP2

During the second 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 and provide new services and features coming directly from the consortium's research. Tagster, the peer-to-peer tagging system developed by the University of Koblenz, is in a very advanced development stage and starts to establish an initial user base from which the system will further evolved. Finally, Ikoru and the zexe.net system, developed by the SONY-CSL team, aim to explore novel ways and contexts for social tagging.

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 Sony 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 available the research done at TAGora 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 artisitic community, and take it even further by introducing tagging to new users and novel contexts.

3.5.2 Progress

In this section we describe the progresses achieved during the second 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 breakthroughs achieved within the project. The applications and products played a significant role in dissemination as expected. The TAGora website was maintained during this second year, and its contents were enriched with news and publications. The efforts to reach wider audiences and introduce them to tagging were succesful, particularly in achieving the involvement of people did not use tags as means of classification before.

Task 5.2.1 - Explicit dissemination activity

In this second year, breakthroughs in research about tagging have effectively reached diverse audiences, thus fulfilling the dissemination objectives. An important number of articles and papers were published in scientific and non-scientific journals, making the results of our research widely available. These breakthroughs were also presented at major conferences, talks and workshops throughout the world, providing a means to avhieve direct contact with interested public. As an example, members from four of the teams from the TAGora consortium were contributing to the ACM Hypertext 2008 Conference in Pittsburgh, USA.

Regarding web-based dissemination, the TAGora website has been 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, approaching the path toward a portal-like look. Simulators, datasets and new products delivered by the consortium can now be downloaded from the website, along with tutorials explaining the functionality of the most popular social tagging websites available on the World Wide Web. The decision of building a full portal has been postponed to the end of the project, when a full deployment of algorithms and control strategies, which are to be delivered during the project lifetime, will eventually be feasible and reliable.

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. Ikoru, the image and music tag-based navigation application, was succesfully put to use in the context of an installation by photographer Armin Linke. In this installation, which is currently being shown at the Zentrum fr Kunst und Medientechnologie Karlsruhe (ZKM) in Germany and the Institute for Contemporary Art and Thought in Athens, Greece, people were able to create printed books with the photographs of Linke and add titles to these books. These titles are considered as tags for books, which are stored in a database using Ikoru as a back-end. In this way, museum visitors were engaged in tagging in a physical space. References to the TAGora project were included both in the books printed by people, the catalog of the exhibition and the space of the exhibition itself.

The GENEVE*accessible project, which uses the zexe.net system, features the TAGora logo in its front page. The project was shown publicly for the first time in April 2008, and is currently being shown in the Centre d'Art Contemporain de Geneve. Since then, the web page of the project has gotten an average of 1.260 visits per day.

Bibsonomy significantly increased its user base, thanks in part to its growing usefulness and variety of features, which were implemented as a direct result of theoretical research. A reference to the TAGora project in Bibsonomy can be found in the Bibsonomy's about/projects section.

The peer-to-peer tagging application Tagster was disseminated to a selected user group for testing it in a real world setting. The announcement to greater community will follow. However, Tagster is already publicly available for download on the web `http://isweb.uni-koblenz.de/Research/tagster`. A publicly accessible web page for the MyTag application, a cross folksonomy search and recommendation tool, was set up by the University of Koblenz team `http://mytag.uni-koblenz.de/`. MyTag allows users to search accross folksonomy sites YouTube, Flickr and del.icio.us. Additionally, users can create a personal profile on MyTag which is used to tailor the search results to the users personal interests. This application was developed within the context of TAGora, and is being used to disseminate the project through the software implementation of theoretical research. MyTag is being used to track user interests and to gather search requests which can later be used for further analysis.

Task 5.2.3 Contribution of Sony CSL.

The last year Sony CSL organized a symposium and open-house, which was an opportunity to present and demonstrate our work. The event being bi-annual, no symposium was organized this year.

Zexe.net started a new project in Geneva, Switzerland, where it is working with a community of

handicapped people who document the state of the accessibility in the city.
The installation 'Phenotypes/Limited Forms' went in display in the ZKM, Karlsruhe, Germany, and at the 'Selective Knowledge' exhibition in Athens, Greece. The approximate 70000 visitors to the ZKM produced more than 8000 books – a selection of 8 photos tagged by the visitor. The exhibition is Greece yielded 16 reviews, notably one in the International Herald Tribune (09/04/08, 'Artists question objectivity') and 21 annoncements in the Greek press.

Task 5.3 Training activities and outreach
Regarding training activities and outreach, the TAGora team has been involved in the following events:

- 1st International Workshop on Collective Semantics: Collective Intelligence & the Semantic Web (CISWeb 2008) at the 5th European Semantic Web Conference (ESWC-08) `http://www.eswc2008.org`, June 2, 2008, Tenerife, Spain
- General and local chair for the 3rd International Conference on Semantic and Digital Media Technology, 2008 (SAMT-2008), Dec 3-5, 2008, Koblenz, Germany
- Program chair for the 7th International Semantic Web Conference, 2008 (ISWC'08), Oct 26-30, 2008 Karlsruhe, Germany
- General chair, ACM IUI-2008, International Conference on Intelligent User Interfaces, January 13-16, 2008, Maspalomas, Spain
- PC Co-Chair, ODBase-2007, The 6th International Conference on Ontologies, DataBases, and Applications of Semantics, Nov 25-30, 2007, Albufeira, Portugal
- Association for the Advancement of Artificial Intelligence 2008 Spring Symposium: Social Information Processing (AAAI-SIP 2008), March 26-28, Stanford University, California, USA
- Symposium "Createurs Singuliers" associated to the GENEVE*accessible project `http://www.zexe.net/GENEVE/geneve.php?c=275`, May 27, 2008, Geneva, Switzerland

3.5.3 Deviations and Corrective Actions

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

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

3.5.4 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
5.4	(Task 5.2) Portal focused on collaborative social systems addressed not only to experts from social, sciences information society, statistical physics but also to a general audience on the web (Month 23).	5	31 May 2008	31 May	2	2	**PHYS-SAPIENZA**

3.6 Workpackage 6 (WP6) - Management

3.6.1 Objectives

Task 6.1. Management

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.

3.6.2 Progress

Task 6.1 Management

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 second 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 (D6.3).

Progress (Milestones)

M6.4 Co-ordination and Management Meetings (Month 23).

Two Project meetings have been organized during the second year.

IV TAGora meeting Kassel, October 25/26 2007;

V TAGora meeting Torino, May 7/8 2008;

3.6.3 Deviations and Corrective Actions

PHYS-SAPIENZA: none

3.6.4 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
6.3	Yearly Management Report (month 23).	6	31 May 2008	31 May	1.5	1.5	**PHYS-SAPIENZA**

Mil. No.	Milestone name	WP No.	Date due	Actual/ Forecast delivery date	Lead contractor
M6.4	Co-ordination and Management Meetings (month 23).	6	31 May 2008	31 May 2008	**PHYS-SAPIENZA**

Chapter 4

Consortium Management

4.1 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:

- **IV TAGora meeting** Kassel, October 25/26 2007;
- PHYS-SAPIENZA and SONY-CSL organized a bilateral meeting in Torino, February 4-7 2008, focused on WP3 and WP4.
- PHYS-SAPIENZA and UNI-SOTON organized a bilateral meeting in Southampton, March 25-28 2008, focused on WP3.
- **V TAGora meeting** Torino, May 7/8 2008;

Frequent contacts among participants were also maintained by e-mail, telephone, occasional visits, short and long term visits.

4.2 Problems, deviations and corrective actions

PHYS-SAPIENZA: 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 we couldn't get the money back from the department interested by the offsetting procedure. Since this is problem internal to "Sapienza" University, we, as PHYS-SAPIENZA team, decided not to propagate the problem to the other teams and we transferred the due budget to all the other teams as expected for the second year. As a result PHYS-SAPIENZA budget for the second year has been greatly reduced. This did not affect considerably the research activity because our department has been able to anticipate part of the money not transferred by the Commission but unless a solution will be found in a very short time, PHYS-SAPIENZA will not be able to keep the same personnel for the third year, reducing in this way the expected results.

series SONY-CSL: none

series UNIK: The work is progressing according to schedule. Because of hiring younger staff than expected, we will need more time (in terms of person months) to fulfill the objectives of the

proposal. However, this will not influence the required budget, due to their comparatively lower salaries.

series UNIKO-KD: none

series UNI-SOTON: Work on all WPs and tasks have been very successful. The only delay is with respect to the development of a semantic recommender system, which we planned to have a prototype developed by month 23. This is now slightly pushed back to month 25 because of the additional time and effort it took to investigate the necessary issues and generate the semantic underlay that this recommender system will be built on. To remedy the problem, additional person months have been secured to the project at no extra cost to speed up the development of this system over the summer of 2008. We are very confident that the system will be up and running in a few months and will be a very valuable resource for testing and disseminating our work on cross-folksonomy analysis and recommendations.

4.3 Project Timetable and Status

Overall the project is progressing as planned and there have been significant progresses in all the Workpackages, in some cases some Workpackages are in significant advance compared to the objectives originally stated.

Workpackages - Plan and Status Barchart (Second year)

Months	12	13	14	15	16	17	18	19	20	21	22	23
WP1 - Emergent Metadata												D1.2 (Task 1.2) D1.2 (Task 1.3)
WP2 - Applications												D2.3 (Task 2.1, 2.2, 2.3) M2.2 M2.3
WP3 - Data Analysis of emergent properties												D3.2 D3.3 D3.5 (Task 3.5)
WP4 - Modeling and simulations						M4.1 (Task 4.1) M4.2 (Task 4.2)						D4.2 (Task 4.1) D4.2 (Task 4.1) D4.3 (Task 4.1) M4.3 (Task 4.1) M4.4 (Task 4.2)
WP5 - Dissemination and exploitation												D5.4
WP6 - Management												D6.3 M6.4

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

Chapter 5

Other issues

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

As partners both of TAGora and the IP ECAgents, PHYS-SAPIENZA and SONY-CSL organized several intiatives, in particular the *International School on Complexity: Course on Statistical Physics of Social Dynamics: Opinions, Semiotic Dynamics, and Language*, jointly with Ettore Majorana Foundation and Center For Scientific Culture, Erice, Italy 13-20 July, 2007 (`http://pil.phys.uniroma1.it/erice2007`). Ciro Cattuto (PHYS-SAPIENZA) organized a satellite workshop of the European Conference on Complex Systems 2007 (ECCS 2007) on *Social websites: complex dynamics and structure*, Dresden, Germany, 5th October 2007. Vittorio Loreto (PHYS-SAPIENZA) took part to the Workshop on the results of the 'Simulating Emergent properties of Complex Systems' pro-active initiative, Brussels, 25th October 2007. Vittorio Loreto (PHYS-SAPIENZA) presented the TAGora project to the Showcase of European Complexity Science Projects (CRP Forum), October the 6th, 2007, Dresden, Germany. Vittorio Loreto (PHYS-SAPIENZA) presented a position paper *Information dynamics in web-based social systems* at the FET COSI-ICT Workshop, "Science of Complex Systems for Socially Intelligent ICT", October the 6th, 2007, Dresden, Germany. Vittorio Loreto (PHYS-SAPIENZA) will be the session leader of the Information and Communication Technologies session at European Conference on Complex Systems 2008 (ECCS 2008), Jerusalem, Israel, September 14th-19th, 2008.

Bibliography

Rabeeh Abbasi and Steffen Staab. Introducing triple play for improved resource retrieval in collaborative tagging systems. In *In: Proc. of ECIR'08 Workshop on Exploiting Semantic Annotations in Information Retrieval (ESAIR 2008)*, 3 2008. URL http://www.uni-koblenz.de/~abbasi/publications/Abbasi2008ITP.pdf.

Rabeeh Abbasi, Steffen Staab, and Philipp Cimiano. Organizing resources on tagging systems using t-org. In *Proceedings of the Workshop "Bridging the Gap between Semantic Web and Web 2.0" at ESWC 2007*, June 2007. URL http://www.uni-koblenz.de/~abbasi/publications/T-ORG.pdf.

Rabeeh Abbasi, Sergey Chernov, Wolfgang Nejdl, Raluca Paiu, and Steffen Staab. Exploiting Flickr Social Information for Finding Landmark Photos. *Submitted in Proc. CIKM*, 2008.

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.

A. Capocci, A. Baldassarri, V.D.P. Servedio, and V. Loreto. Statistical properties of interarrival times distribution in social tagging systems. 2008. to be submitted.

Claudio Castellano, Santo Fortunato, and Vittorio Loreto. Statistical physics of social dynamics, Oct 2007. URL http://arxiv.org/abs/0710.3256. in press in Rev. Mod. Phys.

C. Cattuto, A. Baldassarri, V. D. P. Servedio, and V. Loreto. Vocabulary growth in collaborative tagging systems. 2007a. URL http://arxiv.org/abs/0704.3316.

C. Cattuto, A. Barrat, A. Baldassarri, G. Schehr, and V. Loreto. The collective dynamics of social annotation, 2008a.

Ciro Cattuto. Semiotic dynamics in online social communities. *The European Physical Journal C - Particles and Fields*, 46(0):33–37, 2006. URL http://dx.doi.org/10.1140/epjcd/s2006-03-004-4.

Ciro Cattuto, Vittorio Loreto, and Vito D.P. Servedio. A yule-simon process with memory. *Europhysics Letters*, 76(2):208–214, 2006.

Ciro Cattuto, Vittorio Loreto, and Luciano Pietronero. Semiotic dynamics and collaborative tagging. *Proceedings of the National Academy of Sciences United States of America*, 104:1461, 2007b. URL http://xxx.lanl.gov/abs/cs.CY/0605015.

Ciro Cattuto, Christoph Schmitz, Andrea Baldassarri, Vito D. P. Servedio, Vittorio Loreto, Andreas Hotho, Miranda Grahl, and Gerd Stumme. Network properties of folksonomies. *AI Communications Journal, Special Issue on Network Analysis in Natural Sciences and Engineering*, 20 (4):245–262, 2007c. ISSN 0921-7126. URL http://www.kde.cs.uni-kassel.de/stumme/papers/2007/cattuto2007network.pdf.

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 (ECAI2008)*, 7 2008b. URL http://arxiv.org/abs/0805.2045. http://arxiv.org/abs/0805.2045.

Ciro Cattuto, Dominik Benz, Andreas Hotho, and Gerd Stumme. Semantic grounding of tag relatedness in social bookmarking systems, 2008c.

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.

Olaf Görlitz, Sergej Sizov, and Steffen Staab. PINTS: Peer-to-peer infrastructure for tagging systems. In *Proceedings of the Seventh International Workshop on Peer-to-Peer Systems, IPTPS08*, Tampa Bay, USA, February 2008.

H. S. Heaps. *Information Retrieval: Computational and Theoretical Aspects*. Academic Press, Inc., Orlando, FL, USA, 1978. ISBN 0123357500.

Rene Henkes. Warum wer wen kennt. eine themenspezifisch auswertung sozialer netzwerke. Diploma thesis, Universität Koblenz-Landau, March 2008.

Robert Jäschke, Beate Krause, Andreas Hotho, and Gerd Stumme. Logsonomy - a search engine folksonomy. In *Proceedings of the Second International Conference on Weblogs and Social Media(ICWSM 2008)*. AAAI Press, 2008. URL http://www.kde.cs.uni-kassel.de/hotho/pub/2008/Krause2008logsonomy_short.pdf.

J. Kleinberg. The small-world phenomenon: An algorithmic perspective. In *Proceedings of the 32nd ACM Symposium on Theory of Computing*, 2000.

Jon Kleinberg. Bursty and hierarchical structure in streams. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 91–101, New York, NY, USA, 2002. ACM. ISBN 1-58113-567-X.

Beate Krause, Andreas Hotho, and Gerd Stumme. The anti-social tagger - detecting spam in social bookmarking systems. In *Proc. of the Fourth International Workshop on Adversarial Information Retrieval on the Web*, 2008a. URL http://airweb.cse.lehigh.edu/2008/submissions/krause_2008_anti_social_tagger.pdf.

Beate Krause, Andreas Hotho, and Gerd Stumme. A comparison of social bookmarking with traditional search. In Craig Macdonald, Iadh Ounis, Vassilis Plachouras, Ian Ruthven, and Ryen W. White, editors, *Advances in Information Retrieval, 30th European Conference on IR Research, ECIR 2008*, pages 101–113, 2008b.

Benjamin Markines, Ciro Cattuto, Filippo Menczer, Dominik Benz, Andreas Hotho, and Gerd Stumme. Social similarity, 2008.

F. Pachet. Description-based design of monophonic melodies, 2008. Submitted to Computer Music Journal.

F. Pachet and P. Roy. Analytical features: a knowledge-based approach to audio feature generation, 2008a. submitted to Journal of Artificial Intelligence Research.

F. Pachet and P. Roy. Is hit song science a science?, 2008b. Accepted to the International Symposium on Music Information Retrieval (ISMIR).

F. Pachet and P. Roy. Improving multi-class analysis of music titles: A large-scale study, 2008c. Accepted with major changes, to appear in IEEE Transactions on Audio, Speech and Language Processing.

P. Rabbat and F. Pachet. Direct and inverse inference in music databases: How to make a song funk ?, 2008. Accepted to the International Symposium on Music Information Retrieval (ISMIR).

V.D.P. Servedio, A. Baldassarri, A. Capocci, and V. Loreto. Hunting correlations in streams of text. 2008. to be submitted.

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.

Christoph Tempich. *Ontology Engineering and Routing in Distributed Knowledge Management Applications*. PhD thesis, University of Karlsruhe, 2006.