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)

Final Activity Report

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

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

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

Contents

Abstract

Project Number 034721

Project Acronym TAGora

Project Full Title Semiotic dynamics in online social communities

Date of Delivery Contractual: 31/08/2009 Actual: 31/08/2009

Name of the Report Final Activity Report

Version Final

Authors

- Rabeeh (Univ. Koblenz-Landau) abbasi@uni-koblenz.de
- Alani Harith (Univ. Southampton) ha@ecs.soton.ac.uk
- Baldassarri Andrea (Sapienza Univ.) andrea.baldassarri@roma1.infn.it
- Benz Dominic (Univ. Kassel) benz@cs.uni-kassel.de
- Cattuto Ciro (Sapienza Univ.) ciro.cattuto@roma1.infn.it
- Capocci Andrea (Sapienza Univ.) andrea.capocci@gmail.com
- Dellschaft Klaas (Univ. Koblenz-Landau) klaasd@uni-koblenz.de
- Goerlitz Olaf (Univ. Koblenz-Landau) goerlitz@uni-koblenz.de
- Hanappe Peter (Sony CSL) hanappe@csl.sony.fr
- Hotho Andreas (Univ. Kassel) hotho@cs.uni-kassel.de
- Loreto Vittorio (Sapienza Univ.) loreto@roma1.infn.it
- Maisonneuve Nicolas (Sony CSL) maisonneuve@sony.csl.fr
- Servedio Vito D. P. (Sapienza Univ.) vito.servedio@roma1.infn.it
- Staab Steffen (Univ. Koblenz-Landau) staab@uni-koblenz.de
- Steels Luc (Sony CSL) steels@arti.vub.ac.be

- Stumme Gerd (Univ. of Kassel) stumme@cs.uni-kassel.de
- Szomszor Martin (Univ. Southampton) mns03r@ecs.soton.ac.uk

Contact persons

- Alani Harith (Univ. Southampton) ha@ecs.soton.ac.uk
- Loreto Vittorio (Sapienza Univ.) loreto@roma1.infn.it
- Staab Steffen (Univ. Koblenz-Landau) staab@uni-koblenz.de
- Steels Luc (Sony CSL) steels@arti.vub.ac.be
- Stumme Gerd (Univ. of Kassel) stumme@cs.uni-kassel.de

Keywords List

Folksonomy, Tagging, Semantics, Web navigation, Web 2.0, Complex systems.

Chapter 1

Executive summary

1.1 Introduction

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.

TAGora has been a very timely project that took on the challenge raised by the widespread diffusion of access to the Internet and the new modalities of interaction between Web users and the information available online. The new vision of the Web regarded 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 has become in the last few years 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.

Overall, the collaborative character underlying many Web 2.0 applications put 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, provided 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 aimed at understanding and modeling information dynamics in online communities, providing a solid scientific foundation for the emerging field of "Web Science".

TAGora has been articulated in four main areas whose activities were strongly intertwined:

Emergent metadata Collecting actual raw data from existing, live systems, or from brand new collaborative applications. Several online communities were readily accessible over the web: for a selected set of these systems, tools have been developed and deployed to harvest the relevant data, metadata and temporal dynamics, and to store the acquired information in a form amenable for data analysis. On the other hand brand new applications allowed the Consortium to gain unimpeded access to the raw data and ultimately provided an experimental "clean room" platform that has been employed to validate the understanding of metadata emergence, and to experiment innovative control strategies.

Data analysis of emergent properties Quantitative analysis of folksonomies devoted to:

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

Modeling and simulations The objectives of this research area were twofold: (i) develop models that capture the essence of the emergent dynamics and explain how it might arise from the interactions of single agents; (ii) 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

Feedback and control Feeding back the outcome of the previous activities to the applications developed by the Consortium in order to experimentally verify the devised control strategies and demonstrate the technological advantage achieved by the present project.

1.2 Consortium

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

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

Contractor n. 1 - PHYS-SAPIENZA

The PHYS-SAPIENZA team is composed by a world-level research group in the whole area of statistical physics, information theory and complex systems. The PHYS-SAPIENZA team brought into the consortium its research experience in the field of developing new theoretical tools to collect and analyse data, of introducing and studying suitable modeling for complex system in order to understand the role and the importance of the different factors in a system of communicating agents, of constructing theoretical approaches which could provide with different levels of abstraction and a feedback for new experiments and studies.

PHYS-SAPIENZA is the project coordinator and responsible of WP6 (management) as such. It is also responsible of WP4, whose objectives are related to developing realistic models for the systems studied in WP1, where it contributed to data collecting (in a joint effort of the consortium to download consistent snapshots of Delicious and Flickr), and in WP3, where it actively took part in data analysis (network structure of folksonomies, co-occurrence graphs, semantic grounding of tag similarity, tag dictionary growth, bursty tagging activity). The contribution to WP4 consisted mainly in the introduction of the Semantic Walker Model, with which many folksonomy properties, both global (eg. the tag dictionary growth) and local (eg. structure of the tag cooccurrence networks), can be recovered. In WP2, it has been working on the Live Social Semantics applications as well as on the web-based TAGnet application. As for WP5, PHYS-SAPIENZA contributed largely to the

dissemination of TAGora, through publications in international journals, presentations, workshop and events organisations.

Contractor n. 2 - SONY-CSL

Sony CSL is a fundamental research laboratory, founded in 1996 by Prof. Luc Steels. Research at CSL focuses on three areas: personal music experience, self-organising communication systems and sustainability. The team is interested in the combination of collaborative tagging with musical features that are automatically extracted from data. Collaborative tagging is also considered as an important step towards semi-automatic or automatic music and image categorization. Research in sustainability focuses on the participatory sensing systems, an approach to gather and analyse local data through the participation of the public. In this context, Sony CSL is also interested to explore new forms of tagging for real-world and sustainability-related issues.

Sony CSL is the coordinator of WP5 ('Dissemination'). In WP1 ('Data collection') and WP2 ('Application'), Sony CSL contributed to the design and development of applications of collaborative tagging and the gathering of new forms of tagging data from real-world situations. These efforts were aimed at two different contexts: the support of by small real-world communities facing sustainability-related issues (cfr. Zexe.net and NoiseTube) and the exploration of a new types of artistic installations (cfr. Ikoru). In WP3 and WP4, concerned with data analysis, most of the results are related to the analysis of music metadada to try to extract of tag from audio features in musics or generate variation of music. For WP5, by collaborating with artists and targeted communities having sustainability-related issues, Sony CSL reached an audience beyond the communities that are already interested or familiar with tagging. Furthermore, Sony CSL organized a symposium and open-house in 2006, which was a major opportunity to present and demonstrate results of TAGora to the scientific community.

Contractor n. 3 - UNI KO-LD

The main research focus of the ISWeb group at the University of Koblenz-Landau is the semantic web. Hence there is a profound expertise in semantic technologies like semantic data management and reasoning. Within Tagora Koblenz was coordinating work package 3 Data Analysis of Emergent Properties. In the beginning a significant contribution was done to the data collection in WP1. Several aspects of folksonomy network structure and dynamics were investigated in Koblenz. This includes tag stream analysis as well as deriving models which resemble the characteristics of the users tagging behaviour. Moreover, classification of tags and resources as well as other inferencing tasks were done the large data sets gathered from delicious and flickr. Two applications were developed: Tagster an MyTag. Tagster is a peer-to-peer based tagging systems which employs decentralized organisation of the tagging data. MyTag is a cross-folksonomy search system which is used as as test bed for implementing new approaches for tagging data retrieval. in cooperation with Southampton tag sense disambiguation was integrated.

Contractor n. 4 - UNIK

The research unit Knowledge & Data Engineering in the department of Electrical Engineering/Computer Science at the University of Kassel started in April 2004 with the establishment of a donated chair of the Hertie Foundation. Research in the unit focuses on knowledge engineering, in particular on discovering and structuring knowledge, derivation of new knowledge, and communication of the knowledge. The research unit in particular deals with the development of methods and techniques at the junctions of the research areas Knowledge Discovery, Ontologies/Metadata, Semantic Web, Peer to Peer, Formal Concept Analysis as well as visualization and interaction in

order to reach synergies. The research unit is member in the Research Center L3S, Hannover, Germany.

UNIK is the coordinator of WP2, where it contributes with the social bookmarking and publications management system BibSonomy. BibSonomy is providing data sets on a regular, half-yearly basis for researchers within and out of the TAGora project. UNIK was coordinating the crawl of the del.icio.us data set in WP 1. Based on these datasets, UNIK was studying in WPs 3 and 4 methods for analysing structural properties of folksonomies, for community detection, for semantically analysing tag similarities, for spam detection, for tag recommendations and for analysing user behavior. The most prominent measures and algorithms have been implemented in the BibSonomy platform.

Contractor n. 5 - UNI-SOTON

UNI-SOTON is the main leading partner of WP1 Emergent Metadata. The goal of WP1 is to collect raw data about the static and dynamical properties of folksonomies and, based on such data, to identify sylized facts (emergent feautures) for subsequent investigation and modeling. The activity of data collection covers the design and development of software clients, the deployment of clients, and the post-processing of raw data to produce high-quality large-scale datasets to be used for data mining.

Although not in the initial plan of the project, UNI-SOTON made a big contribution to application building in WP2 Applications, by providing several important services to MyTag; a prominent application lead by UNI KO-LD. UNI-SOTON was also the leading partner behind the design, implementation, and deployment of Live Social Semantics (D2.5 and D4.5); an application designed in collaboration with the SocioPatterns.org project to show off some of TAGora's technologies at major international conferences. This application proved to be a great success and highly impacting dissemination activity.

In WP3 Data analysis of emergent properties, UNI-SOTON lead the way towards cross-folksonomy analysis and integration of data. This involved building semantic models for representing such heterogenous and distributed data, designing and deploying various tools for joining such data into single profiles, processing tag clouds for identifying interests of individuals, filtering raw tags into cleaner sets of terms, etc. Many of such tools are currently freely accessible online, providing a valuable service for anyone with interest in tag analysis.

With respect to WP4, UNI-SOTON was responsible for exploring how folksonomies and social tagging could be used to provide recommendations to users. To this end, much work was spent on researching tagging behaviour, syntactical and semantic properties of tags, cross-linking datasets and folksonomies for recommendation purposes, and building several services for recommending tag senses, and user interests. These recommendation services were successfully integrated into MyTag and LSS.

UNI-SOTON contributed largely to the dissemination of TAGora, through publications, presentations, workshop and events organisations.

1.3 List of Deliverables

Del. No.	Deliverable name	WP No.	Date due	Actual/ Forecast delivery date	Estimated indicative person-months	Used indicative person-months	Lead contractor
D1.1	Data delivery from selected folksonomy sites (Month 11).	1	15 Jul 2007	20 June 2007	1	1	**PHYS-SAPIENZA**
D1.2	(Task 1.2) Data delivery from bibliographic reference sharing systems (Month 23).	1	15 Jul 2008	20 June 2008	3	3	**UNI KO-LD UNIK**
	(Task 1.3) Data delivery from experimental tag-based navigation systems (Month 23).	1	15 Jul 2008	20 June 2008	2	2	**SONY-CSL**
D1.3	(Task 1.4) Data delivery from selected recommender systems (Month 11).	1	15 Jul 2007	20 June 2007	4	4	**UNI-SOTON**
D1.4	(Task 1.5) Public delivery of data collected by the Consortium and related documentation (Month 38).	1	15 Oct 2009	15 Sept 2009	4	4	**UNI-SOTON (ALL)**
D2.1	Task 2.1a First version of social tagging system for bibliographic data (Month 11).	2	15 Jul 2007	20 June 2007	4	4	**UNIK**
	Task 2.1b First version of folksonomy peer-to-peer system for sharing of bibliographic data (Month 11).	2	15 Jul 2007	20 June 2007	7.5	7.5	**UNI KO-LD**
D2.2	Task 2.2a First version of the Tag-based navigation system for images (Month 11).	2	15 Jul 2007	20 June 2007	3.5	3.5	**SONY-CSL**
	Task 2.2b First version of the Tag-based navigation system for music (Month 11).	2	15 Jul 2007	20 June 2007	3.5	3.5	**SONY-CSL**
D2.3	(Task 2.1, 2.2., 2.3) Interim report on tagging systems update and usage (Month 23).	2	15 Jul 2008	20 June 2008	2	2	**UNIK (ALL)**

D2.4	(Task 2.2) Final version of the Tag-based navigation system for images (Month 38). (Task 2.2) Final version of the Tag-based navigation system for music (Month 38).	2 2	15 Oct 2009 15 Oct 2009	15 Sept 2009 15 Sept 2009	2 2	2 3	**SONY-CSL** **SONY-CSL**
D2.5	(Task 2.1) Final report on tagging systems update and usage (Month 38).	2	15 Oct 2009	15 Sept 2009	2	2	**UNIK (ALL)**
D3.1	Tools and report for extracting emergent metadata statistics and network metrics. Based on data formats described in WP1, in these tools will provide basic statistical (cf.Task 3.1) and network analysis data (cf. Task 3.2) represented in an agreed and reusable format (Month 11).	3	15 Jul 2007	20 June 2007	5	5	**UNI KO-LD**
D3.2	Methods for identifying communities (Month 23).	3	15 Jul 2008	20 June 2008	1	1	**UNIK (ALL)**
D3.3	Methods for using semantic inference in data analysis (Month 23).	3	15 Jul 2008	20 June 2008	2	2	**UNI KO-LD**
D3.4	Methods for tracking *tag* emergence (Month 38).	3	15 Oct 2009	15 Sept 2009	1	1	**UNI-SOTON**
D3.5	(Task 3.5) Protocol for integrating cross-folksonomy networks (Month 23).	3	15 Jul 2008	20 June 2008	2	2	**UNI-SOTON**
D4.1	(Task 4.1) Review of theoretical tools for modeling and analysing Collaborative Social Tagging Systems (Month 11).	4	15 Jul 2007	20 June 2007	2	2	**PHYS-SAPIENZA**

D4.2	(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 June 2008	2	2	**PHYS-SAPIENZA (ALL)**
	(Task 4.1) Report on the roadmap leading from modeling activity to control strategies (Month 23).	4	15 Jul 2008	20 June 2008			**PHYS-SAPIENZA (ALL)**
D4.3	(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 June 2008	2	2	**PHYS-SAPIENZA**
D4.4	(Task 4.2) Review of existing recommendation strategies and systems (Month 11).	4	15 Jul 2007	20 June 2007	2	2	**UNI-SOTON**
D4.5	(Task 4.2) Deployment of a semantic recommender (Month 38).	4	15 Oct 2009	15 Sept 2009	3	3	**UNI KO-LD**
D4.6	(Task 4.2) Report describing the results of the control experiments performed (Month 38).	4	15 Oct 2009	15 Sept 2009	2	2	**PHYS-SAPIENZA (ALL)**
D5.1	(Task 5.1) Project presentation report (Month 4).	5	30 Sept 2006	30 Sept 2006	1.5	1.5	**PHYS-SAPIENZA**
D5.2	(Task 5.2) Website for the project (Month 4).	5	30 Sept 2006	30 Sept 2006	2	2	**PHYS-SAPIENZA**
D5.3	A *White Paper* that will describe target problems and grand challenges for Semiotic Dynamics Systems, clearly recognized by all and openly communicated to the scientific community (Month 11).	5	15 Jul 2007	20 June 2007	2.5	2.5	**PHYS-SAPIENZA (ALL)**
D5.4	(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	15 Jul 2008	20 June 2008	2	2	**PHYS-SAPIENZA**

D5.5	(Task 5.3) Report on the impact, usability and user communities characterization of our web-based experiments and demos (Month 38).	5	15 Oct 2009	15 Sept 2009	2	2	**SONY-CSL**
D6.1	Provision of reports as required to the Commission.	6	15 Jul 2007	20 June 2007	1	1	**PHYS-SAPIENZA**
D6.2	Yearly Management Report (month 11).		15 Jul 2007	20 June 2007	2	2	**PHYS-SAPIENZA**
D6.3	Yearly Management Report (month 23).	6	15 Jul 2008	20 June 2008	1.5	1.5	**PHYS-SAPIENZA**
D6.4	Yearly Management Report (Month 38).	6	15 Oct 2009	15 Sept 2009	4	4	**PHYS-SAPIENZA**

1.4 List of Milestones

Mil. No.	Milestone name	WP No.	Date due	Actual/ Forecast delivery date	Lead contractor
M1.1	(Task 1.1) Implementation of software clients and hardware infrastructure to perform data collection from folksonomy sites (Month 5).	1	31 Oct 2006	31 Oct 2006	**PHYS-SAPIENZA SONY-CSL UNI KO-LD UNIK**
M1.2	(Task 1.4) Design and deploy a centralised system for storing the selected online resources (Month 5).	1	31 Oct 2006	31 Oct 2006	**UNI-SOTON**
M2.1	First version of social tagging system for bibliographic data (Month 5).	2	31 Oct 2006	31 Oct 2006	**UNIK**
M2.2	Definition of the control strategy and decision about improvements for the final version of the system for images (Month 23).	2	31 May 2008	31 May 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	31 May 2008	31 May 2008	**SONY-CSL**
M3.1	Acquisition of software tools for data analysis (Month 5).	3	31 Oct 2006	31 Oct 2006	**UNI KO-LD**
M3.2	Identification of the key emergent features and global observables relevant for modeling (Month 5).	3	31 Oct 2006	31 Oct 2006	**PHYS-SAPIENZA**
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	**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	31 May 2008	**UNI-SOTON**
M4.5	(Task 4.2) Preliminary control experiments performed (Month 32).	5	28 Feb 2009	28 Feb 2009	**PHYS-SAPIENZA**

M5.1	Identification of third parties (SMEs) suitable for the deployment of the web-based part of the dissemination plan (Month 11).	5	31 May 2007	31 May 2007	**SONY-CSL**
M5.2	The Sony CSL biannual public symposia (2006) in Paris.	5 2006	31 Oct 2006	6 Oct	**SONY-CSL**
M5.3	The Sony CSL biannual public symposia (2008) in Paris (Month 32).	5	28 Feb 2009	cancelled	**SONY-CSL**
M6.1	Set up of the project information infrastructure (WWW pages, mailing list, ftp area etc.) (Month 3).	6 6	31 Aug 2006	31 Aug 2006	**PHYS-SAPIENZA**
M6.2	Co-ordination and Management Meetings (month 0).	6	30 June 2006	30 June	**PHYS-SAPIENZA**
M6.3	Co-ordination and Management Meetings (month 11).	6	31 May 2007	31 May	**PHYS-SAPIENZA**
M6.4	Co-ordination and Management Meetings (month 23).	6	31 May 2008	x	**PHYS-SAPIENZA**
M6.5	Co-ordination and Management Meetings (Month 38).	6	31 Aug 2009	31 Aug	**PHYS-SAPIENZA**

Chapter 2

Project execution

2.1 Workpackage 1 (WP1) - Emergent Metadata

2.1.1 Objectives

The goal of WP1 is to collect raw data about the static and dynamical proper ties of folksonomies and, based on such data, to identify sylized facts (emergent feautures) for subsequent investigation and modeling. The activity of data collection covers the design and development of software clients, the deployment of clients, and the post-processing of raw data to produce high-quality large-scale datasets to be used for data mining.

Task 1.1 Data from collaborative tagging (folksonomies)

The goal of this task is to provide the project with large-scale and well documented datasets from existing folksonomies. These data became the foundation for an extensive scientific investigation of the statistical properties of folksonomies, as well as used to gain insight into user behavior and tagging patters. The objectives of this task comprised the identification of suitable data sources, the development and deployment of software clients for data collection, the post-processing of data, and the assembly of data repositories for the project, and related documentation.

Task 1.2 Data collection from the bibliographic reference sharing system BibSonomy

BibSonomy is service introduced by UNIK, for managing and sharing web pages User of the open access system can centrally store bookmarks for web pages in form of URLs and add tags to those resources. This task was responsible for collecting BibSonomy data and sharing it with the other project partners. This data was extensively used in research community evolution, tag recommendation, and spam tagging.

Task 1.3 Data from experimental tag-based navigation system at SONY-CSL

SONY-CSL developed several tagging application, for music, images, noise reports, etc. This task aimed at gathering data from these deployments and use them as sources for the analysis tasks to be carried out by members of the TAGora project.

Task 1.4 Collecting data from online recommendation systems

Many e-commerce systems that exist on the web provide some sort of recommendations. Similarly, many folksonomies provide recommendations of tags, resources, groups, etc. This task is concerned with collecting data from such source (e.g. LastFM, Delicious, Flickr) to support our research into semantic recommendation tools and services.

2.1.2 Contractors Involved

Data gathering was a major activity, especially in the first year of the project. **All the partners** collaborated extensively to collect large amounts of data from Delicious, Flickr, and several other smaller sites.

2.1.3 Work Performed

The consortium started to collaborate to gather data very actively from the very beginning of the project. This was fueled by a realization of the necessity of owning such datasets for performing the tasks outlined in the project. To this end, many large and valuable datasets were collected relatively quickly. Most of those datasets where made public through the TAGora website. In some cases, the scripts to gather the data were also published.
The most significant datasets collected by TAGora include:

- **Delicious tags and tagged resources**: Data from Delicious was gathered in 2006 and currently consists of over 667 thousand users, nearly 2.5 million tags (organized in 667 bundles), and around 18.7 million resources. This data set was extensively used in the project, for example in the analysis and modelling of evolutional behaviour and structural information of social resource sharing systems, analysis and modelling of the structure and dynamics of folksonomies, and in semantic user interest profiling and tag disambiguation analysis.

- **Flickr photos:** This data collection contains all descriptions of photos that were uploaded to Flickr during January 2004 and December 2005 and that were still available over the public API in the first half of 2007. The crawling of the data collection was finished 07/2007. The collection contains information about 320K users, 28M photos, 1.6M tags and 113M tag assignments.

- **BibSonomy:** To provide the Consortium with raw data for modeling and analyzing interactions in online social communities, we offer a benchmark dataset from our collaborative tagging system BibSonomy. The anonymized data of BibSonomy are downloadable via a MySQL dump, which will be updated every half year. Interested people get an account from kde@cs.uni-kassel.de for access to our server on https://www.kde.cs.uni-kassel.de/bibsonomy/dumps. Before starting the download, participants have to sign a license agreement in which terms of use are set up. The data set currently consists of over 2.6 thousand users, 181 thousand bookmarks, 219 thousand publications, and over 816 thousand tag assignments. The dumps can easily be loaded into a MySQL database.

- **Zexet.net - MOTOBOY:** The dataset from the canal*MOTOBOY project[1], 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 São Paulo, Brazil transmitted 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 over 8000 tag assignments, nearly 8000 resources, 712 tags and 15 users.

- **"Phenotypes/Limited Forms":** "Phenotypes/Limited Forms" is an art installation that uses photos by the photographer Armin Linke and that has been on display at the exhibitions in Karlsruhe and Siegen (Germany), Athens (Greece) and in São Paulo (Brazil). We collected data about 24000 users, 2400 photos, 17000 tags, and 190000 tag assignments.

- **Tag Senses:** The TAGora Sense Repositor (TSR) is a linked data enabled service endpoint that provides extensive metadata about tags and their possible senses. When the TSR is queried with a particular tag string, by forming a URI that contains the tag in a REST style, the tag is processed, grounded to a set of DBPedia.org resources, and an RDF document is returned containing the results. Creating this dataset involved processing the XML dump of all Wikipedia pages and indexing all titles, then mining redirection and disambiguation links,

[1] http://www.zexe.net/SAOPAULO

then extracting term frequencies for each page. The TSR services that is based on this dataset is currently publicly accessibly from the TAGora site. This dataset currently consist of over 51 thousand processed tags, over 197 thousand sense matchings, and nearly 6.5 million senses from DBpedia processing. This RDF dataset now contains well over 160 million triples.

2.1.4 Final Results

A description of each of the main datasets collected and used by TAGora is available on www.tagora-project.eu/data. Links to where some of those datasets can be downloaded are also given on that site, along with information on what the dataset consists of, size, format, etc.

Several of those datasets have been converted into RDF and made available for download, to facilitate reuse of the data. For Delicious and Flickr, scripts for gathering tagging information were also shared on the project website, to help interested users in collecting their data in RDF.

In addition to making the datasets and crawling scripts publicly available, we have also made several tools and services freely accessible. This includes data analysis tools and tag processing services.

2.2 Workpackage 2 (WP2) - Applications

2.2.1 Objectives

Collaborative tagging originated from the need to manage large collections of data. Tagging data is a means to describe, search, and retrieve objects in an intuitive way, which constitutes an important factor of its success. The objective of this work-package was twofold. It provided 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. The first objective involved building systems that add value to existing tagging sites. One possibility is to enrich navigation based on tags by adding data analysis. The combination of data features and tagging allows to overcome shortcomings of tag-based search, such as problems caused by synonymy, homonymy, missing tags, or spelling mistakes. The added value of our systems was important in order to attract users and thus fulfil our second objective: to serve as a valuable source for data delivery of WP1. Our systems allow us to gain unimpeded access to the raw data and provide an experimental “clean room” platform that was employed to validate our understanding of metadata emergence, and to experiment with the control strategies devised in WP4.

The objective of this work-package was to provide the Consortium with experimental platforms for data collection and for evaluating selected algorithms. In order to have privileged and controllable data sources for the collaboration, we designed and deployed systems — both online systems and actual demonstrations/experiments — for the specific purpose of data collection. This allows unimpeded access to the raw data and will ultimately provide an experimental "clean room" platform that will be employed to validate our understanding of metadata emergence, and to experiment with the control strategies devised in WP4. As such, the results of WP2 are of a technological nature and benefit a broader community than the Consortium itself.

2.2.2 Contractors Involved

UNIK

UNIK is running *BibSonomy*, which users collaborative organizing and sharing of bookmark collections and publication lists. A basic version of BibSonomy has been online before the start of the project. Within Tagora, UNIK has extended its functionality to attract a significant number of users, and have provided means for a systematic generation of data for experiments.

UNI KO-LD

UNI KO-LD has been working on the two systems Tagster and MyTag. *Tagster* is a system for collaboratively organizing and sharing multimedia data in a peer-to-peer network. It is completely decentralized and provides the same functionalities as common centralized folksonomy systems like Flickr, delicious or BibSonomy. The *MyTag* system aims at solving the limitations of current tagging platforms by enabling cross-media search across images, video, and social bookmarks. It offers transparent access to different single-media platforms currently including Flickr, YouTube, and del.icio.us.

SONY-CSL

SONY-CSL has been working on the systems Ikoru, NoiseTube and Zexe.net. The *Ikoru* system is primarily used to experiment with collaborative tagging and content-based analysis. The project consists of a server-side component and a Web interface. The *NoiseTube* platform is an extension of the *Zexe.net* platform which was aimed at finding new ways to use collaborative tagging for offline communities facing issues related to accessibility and sustainability. NoiseTube extends this idea to new type of resource: the exposure of individual citizens to pollution.

PHYS-SAPIENZA

PHYS-SAPIENZA has been working in collaboration with the ISI Foundation (Turin) on *TAGnet*, a prototype (http://www.netr.it) designed to provide users with a reflexive tool to expose regularities and patterns in their own tag-based annotations. Tagging patterns can reveal a lot about a user's experience, her interests and her emergent conceptualizations, but users are not aware of these patterns until these regularities are made explicit by means of data analysis and state-of-the-art visualization. TAGnet currently focuses on Flickr users, providing them with a "semantic mirror". This application was not initially foreseen, and was set up to exploit the results of WP3 and WP4 on the structure of tag co-occurrence networks.

UNI-SOTON

UNI-SOTON, in collaboration with PHYS-SAPIENZA (through the ISI Foundation) and with the SocioPatterns.org project, has developed the the Live Social Semantics system. It illustrates possibilities of utilising various TAGora technologies for the analysis of the social behaviour of conference participants. It was deployed at two major conferences in 2009. This application was not initially foreseen, and was set up to exploit the results of WP3 and WP4 on the social behavior of conference participants.

2.2.3 Work Performed

As the circle from applications to models and back to the applications was central to the project, all contractors were working on implementations. Here, we briefly describe the steps deployed, before presenting the implemented components in more detail in the following subsection 'Final Results'.

BibSonomy – a social resource sharing system for bookmarks and publications

A basic version of BibSonomy has been online at the start of the project. Our work in Tagora had three objectives: providing more functionality in order to attract more users, implementing the best-working models of WPs 3 and 4, and implementing logging facilities to observe their usage. The extended functionality was in our focus of the first two project years, while the logging component and the implementation of the models were tasks of the second and third project year.

Tagster – Folksonomy Peer-to-Peer System for Sharing Multimedia Data

Tagster is a peer-to-peer tagging application without central storage of tagging data. In the first year, we developed a prototype with basic functionalities for tagging multimedia data in a decentralized fashion. During the second year the application was extended in two directions: the management of distributed tagging statistics and improvements concerning the user interaction and usability. Due to technical problems, limited resources, and difficulties to attract a critical mass of users the work on Tagster was discontinued in the third year.

Ikoru – A Test-bed for Collaborative Tagging and Content-Based Analysis

The Ikoru system, is primarily used to experiment with collaborative tagging and content-based analysis. The project consists of a server-side component and a Web interface, which can be viewed at http://www.ikoru.net. We have made the first version of Ikoru available at the end of the first project year. In the second year we have extended the similarity search to audio. We have kept the Web site up and running since last year but we have been focusing more on targeted tagging experiments than on the growth of the website.

Zexe.net – A Community Memory for Representing Daily Experiences using Folksonomies

The Zexe.net initiative focused on finding new ways to use tagging for the benefit of off-line communities facing real-world issues related to accessibility and sustainability. *canal*MOTOBOY* and *GENEVE*accessible* are the latest deployments of the Zexe.net platform. Both of these projects made intensive use of collaborative tagging. In *canal*MOTOBOY*, 15 motorcycle messengers in São Paulo, Brazil, used multimedia mobile phones to capture images and videos of their daily lives, describe the contents using tags and published them on the web. The *GENEVE*accessible* project involved handicapped people in Geneva, Switzerland. They used multimedia phones equipped with GPS receivers to create maps of their city's accessibility. They used tags to describe the images of obstacles they find in their way. By publishing these (geo-)tagged images on the web, they effectively built an intelligent, collaborative map which is immediately available to the public. These projects have enabled members of TAGora to study the dynamics of tagging in small-scale groups with shared interests. In particular, these projects provide two contrasting examples. While *canal*MOTOBOY* was totally open-ended, *GENEVE*accessible* had very specific goals. We have studied how these different scopes affect the projects' folksonomies.

NoiseTube – Community Memory representing the exposure of people to noise pollution

NoiseTube is the natural successor of the Zexe.net platform. During this 3rd year of TAGora, the application of collaborative tagging was extended to new type of resource: the exposure of individual citizens to pollution. In the NoiseTube project the concrete focus was put on the problem of noise pollution. A platform was created that support the sensing (as in measuring) and tagging of occurrences of noise pollution in cities using the mobiles phones of citizens. The project consists of a web portal (which can be visited at http://www.noisetube.net) and an application for mobile phones.

TAGnet – a tool for awareness and management of personal metadata

TAGnet is a prototype web-based application (http://www.netr.it) designed to provide users with a reflexive tool to expose regularities and patterns in their own tag-based annotations. Tagging patterns can reveal a lot about a user's experience, her interests and her emergent conceptualizations, but users are not aware of these patterns until these regularities are made explicit by means of data analysis and state-of-the-art visualization. TAGnet currently targets Flickr users and provides them with a view on their annotations (tags) that exposes actionable information.

MyTag – a tool for integrating folksonomies

MyTag is a cross folksonomy search tool which allows for searching different resource types like photos, videos and bookmarks from different folksonomy sites in parallel. The first prototype included Flickr, Delicious and YouTube. Additional folksonomy sites like Bibsonomy and Connotea were added subsequently. MyTag includes a personalized ranking and a search assistant that supports the disambiguation of entered search terms.

Live Social Semantics

This application was fully developed and deployed at two major conferences in 2009, ESWC and HT. At both conferences, attendees were able to register to use this application, which provided them with various services aiming at encouraging social interactions. Hundreds of users have already registered, and we are currently in the process of redesigning the application to make is

constantly available on the web. Users of this application will be able to carry forward their data and profiles to any event where this application will be running.

2.2.4 Final Results

BibSonomy – a social resource sharing system for bookmarks and publications

Extended functionality The following features have been implemented in BibSonomy in the Tagora project. These features were announced on a weekly basis on http://bibsonomy.blogspot.com/.

- BibSonomy contains more than 100 explicitly defined groups, which actively sharing and collecting resources. The groups are mostly research groups or participants of European projects.
- To find content from users who used similar tags as oneself, BibSonomy supports more user flexibility regarding different search mechanisms. The following search strategies were implemented after 1st of June 2007 and are available online for navigation: searching by author, searching by concepts, ranking.
- BibSonomy has been integrated with the following third party products: CiteSmart, Citavi, Typo3, Zope, XWiki, WordPress, Moodle, Zotero, OpenID, and some online catalogues of university libraries. There exists now also a firefox addon 'GoogleSonomy', which allows to search in parallel both in Google and in BibSonomy.
- We have implemented an application programming interface (API) which allows external applications to interact with BibSonomy. One example application which uses the API to access BibSonomies data is the stand alone BibTeX Manager JabRef.
- Scrapers for automatically extracting references from over 60 digital libraries or publishers' websites have been provided. The scraping service can be used independently from BibSonomy for other purposes.
- As researchers and students of different nationalities work with BibSonomy, version 2.0 supports multilingual pages. Almost all non-posting pages of BibSonomy are now available in English and German.

Figure 2.1 shows that these features have indeed attracted a significant number of active users.

Implementation of the results of WPs 3 and 4 We have implemented two frameworks within BibSonomy, one for tag recommendations, and one for spam detection. The *recommendation framework* allows integration and evaluation of different recommender systems into BibSonomy. These recommender systems can be either installed locally or remotely (connected and queried via http), thus allowing other research teams to integrate their recommender systems and giving a broad base for evaluation. The framework's central component is a multiplexer where each tag recommender system is registered during initialization. Whenever a user wants to assign tags to a resource, each recommender is queried for recommended tags in parallel, spawning separate threads for each recommender query. All responses are collected and exactly one recommender's result is uniform randomly chosen and presented to the user. Finally, to allow machine learning techniques, we pass to each recommender the set of tags which the user assigned to the resource. We also capture those situations, where the user is unsatisfied with a set of recommended tags: A 'reload' button is displayed, which replaces the set of recommended tags with a different recommender's suggestion. This encourages the user to give us the desired feedback, as new tags are

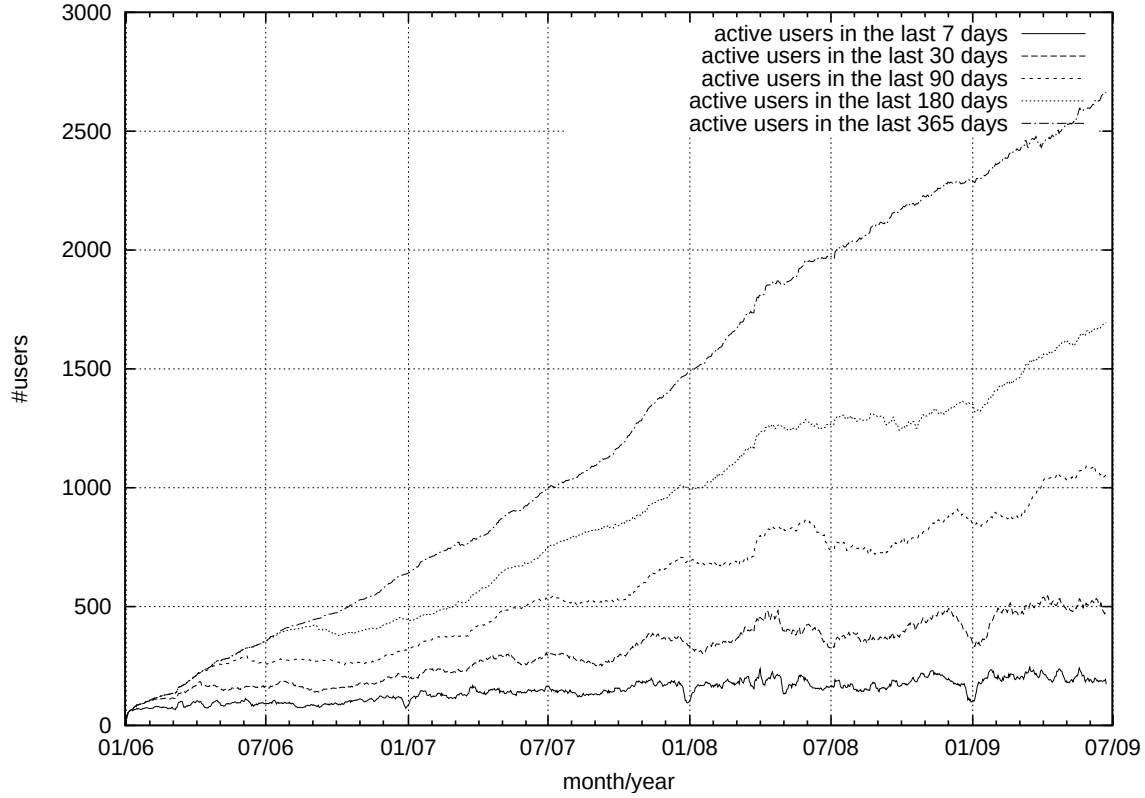


Figure 2.1: Growth of BibSonomy

presented with low cognitive effort. Because of the limited budget of Tagora, only the design of the recommendation framework could be done within the project. Its implementation was done during the extended life time of the project – and used in the ECML PKDD discovery challenge 2009[2] for dissemination of Tagora (see below) – but financed by a national follow-up project.

BibSonomy's *spam framework* has been designed to automatically detect spam posts, and prevent those posts from being viewed by legal users or crawled by search engines. The framework has been developed using the insights we obtained from our first experiments with the BibSonomy Spam dataset.

On a regular basis, the system selects the classified users from the last months. We do not use the entire dataset as spamming behaviour changes over time. The system computes different user features (location, activity and profile based) to create a training dataset. A model is learned from this dataset. The model is serialized and stored. It can be reused for the classification task. The classification task starts every three minutes. New users with at least one public post are collected from the database. A white-list checks, if the user's mail addresses or the IP addresses come from a well known university. Those users are directly marked as non-spammers. The remaining users are classified, using a model of the training phase. Four different categories for classification are available. The classifier predicts spam and non-spam with a certain confidence. Users with a label of one of the three categories spam, spam not sure, spam sure, are marked as spammers. As a consequence, public posts of those users are set to spam posts and can only be viewed by the spammers themselves. The classification results are presented in an admin spam interface. Administrators can go through the lists of classifications and change the decision of the classifier. The implemented spam admin interface can also be used to manually create spam datasets for research purposes.

[2] http://www.kde.cs.uni-kassel.de/ws/dc09/online

Logging User Interaction ClickLog is an add-on for bibsonomy to detect and log user interaction on every bibsonomy web page. Every time a user clicks on an anchor within a page, Clicklog recognises it and computes different values which it submits to a log server. The attributes of a clicked hyper link, its containing text, and its position within the text are used to get values like anchor title, hyperreference and some id and class values to determine the area of this clicked link. Anchor tags in different areas like navigation, tag cloud, bookmark posts, and publication posts have different class definitions. Next, the position of the anchor in a list of anchors will be determined for later analysis. Further meta data that are logged include date, current webpage location, user agent of the web client, username, session id, complete http request header and at least the mouse position while clicking the anchor in two xy-value-pairs, mouse position in client window and position in document area, which are different, if the user scrolls the document.

The click log data can be used for analysing the usage behavior of the different users. Typical questions are: Are there types of links that are used below expectation? Are there specific user pages or tag pages that are accessed above average? Are they accessed from the general tag cloud or from some content-specific pages? How intensively are the relations used? Will new features be used, or are they ignored? How many external links are followed, and where to? How is the usage distributed over the two categories ob bookmarks and publications?

Selected Dissemination Activities We were organising the Discovery Challenge of the ECML/PKDD 2008 conference.[3] The challenge comprises two tasks: learning tag recommendations, and detecting spam, both based on a BibSonomy dataset. The final results were published in September 2008. Because of its success, the conference organizers asked us to repeat the challenge in 2009; its preparation is currently ongoing. Both challenges were organised as Tagora outreach.

Tagster – Folksonomy Peer-to-Peer System for Sharing Multimedia Data

Tagster is a peer-to-peer tagging application. Very much like Flickr, Del.icio.us, 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. Tagster is based on a modular architecture which provides the basic functionality for organizing and sharing annotated information resources in a decentralized scenario. Additionally, a mechanism for managing distributed tagging statistics is integrated and the application provides different data views for easily navigating the annotated multimedia data. Tagster has been published as open source on Launchpad (https://launchpad.net/tagster).

Distributed Statistics

To make tagging meta data available to all users in the network, Tagster uses a global index structure. That means each peer in the network stores a fraction of the globally available meta data and the underlying index implementation (we use Bamboo[4]) assures that the stored data can be accessed in a very efficient way. However, the index only stores pairwise relations between users, tags, and resources. Handling more complex information retrieval tasks like similarity computation of users and result ranking would require contacting many peers to gather the necessary information which is apparently highly inefficient. Therefore, we have developed a novel mechanisms for managing distributed statistics, called PINTS Görlitz et al. (2008).

The basic idea for distributed meta data management is that each peer in the network is maintaining a fraction of the global meta data. With each new tag assignment the responsible index peer

[3] http://www.kde.cs.uni-kassel.de/ws/rsdc08/
[4] http://bamboo-dht.org

updates the respective index information and notifies other peer about the changes if necessary. The similarity computations are based on the cosine similarity of feature vectors, as for example tag clouds. We adapted the well known TF-IDF measure from information retrieval such that each feature combines local and global data like a user's tag frequency and the tag's popularity in the whole network. The problem, however, is to keep the statistics accurate since the propagation of every change of the global index data would cause a high message complexity. Therefore, the PINTS algorithm only propagates data updates if the estimation of the change in the depending statistics is higher than a certain threshold. That allows us to maintain accurate distributed statistical information while keeping the message complexity low in the network. PINTS is implemented in Tagster and used to display statistical information like tag clouds.

User Interface

The design of the user interface plays a major role for the usability of a software. In the case of Tagster it is important to have an interface that provides the same functionalities like the centralized folksonomy systems but also includes an intuitive way of navigating though both the personal data on the local machine and the information retrieved from the network.

However, the application's appearance is not the only aspect we consider for a good user interface but also the ease of use, i.e. the simplicity of configuration and setup/joining of the peer-to-peer network.

Navigation elements The adaption of typical navigation elements like tag clouds from the centralized systems is strongly motivated by the fact that user of such systems are already very accustomed to that type of navigation support. Therefore, one goal is to integrate the same or similar data representations such that users can get familiar with Tagster really quick. This includes, for example, the display of related information for the currently selected data items and contextual tag clouds which are a very typical navigation element.

Additionally, we have integrated resource-specific type views, i.e. the user can browse the resources by their associated Mime type. Thus, it is possible to filter a search results such that only images or documents are displayed.

Tagster's local resource organizations allows the user to tag any file on the local harddisk. Thus, also a file's path information is preserved and displayed in the resource view. To better visualize the local resources we have implemented a tree view that orders all local file in their actual folder hierarchy. Resources from the network are displayed without a hierarchy since that information is not returned for privacy reasons.

Additional functionality Since browsing of resources in the network is not enough, we have also implemented a download protocol for directly retrieving files from other users. To download a resource from the network the user just has to click on the download button next to the resources displayed tagging data. Then the file will be retrieved from the owner and saved in the local download folder. All tags already assigned to the resource will be automatically applied to the downloaded file, too.

Tagster's resource view only diplays the typical file information like name and path. To actually see the content of the files we integrated a function to start the appropriate external application that is associated with the respective file. The intention is that the user does not need to switch manually to another application to view his resources. Currently, this function is supported on Windows and Gnome-based Linux systems.

Ikoru – A Test-bed for Collaborative Tagging and Content-Based Analysis

At the outset of the project, we nourished the hope that Ikoru could grow into an active website. Despite the strengths of the system, this was somewhat wishful thinking. The reality is that in the last two years many sites have integrated tagging and that these sites can rely on considerable resources and infrastructure to continuously improve their offering. Technology transfers within Sony have been in principle possible and Ikoru has been presented to many product division within the group. However, the collaborations have been not trivial to set up because of the current tendency of Sony to outsource Web services.

As a result Sony CSL does not have a precise planning to promote Ikoru to a large audience. Instead, the current strategy is to continue to increase the usability and reliability of the software through its use in small but concrete projects. These focused projects can be managed much more easily and allow us to concentrate on innovative applications of tagging. In the future, we see Ikoru evolve as a generic back-end to store the information about resources, people, and tag assignments. We also see the focus of the tagging applications move away from purely Web-based applications towards real-world applications.

One such project is the artistic installation "Phenotypes/Limited Forms" that was exhibited at the Zentrum für Kunst und Medien (ZKM) in Karlsruhe, Germany, the Bienal de Sao Paulo in São Paulo, Brazil and the "Selective Knowledge" exhibition in Athens, Greece and still is on display in the Museum of Contemporary Art in Siegen, Germany. Although this installation – a joint project with photographer Armin Linke – is a very particular use of Ikoru, it has allowed us to gather a fair amount of data. More than 8000 visitors picked a selection of eight photos and tagged it using a special “editing table” designed for this purpose. The photos, printed in high-quality on solid boards, are taken from an archive of one thousand photos that are displayed on shelves in the exhibition space. Once the visitors tagged their selection, the editing table prints out a small booklet that they can take home.

Another interesting development, that has recently started is the use of Ikoru to store musical melodies (Pachet (2009)). Compared to photos or audio files, melodies can be analysed and generated at a semantically higher level. It has also the potential to reach a small but passionate community.

To facilitate such small tagging projects by other researchers and developers, and to let Ikoru evolve accordingly, we made the source code available under the GNU Library General Public License (LGPL). It can be found at http://sourceforge.net/projects/ikoru.

Zexe.net – A Community Memory for Representing Daily Experiences using Folksonomies

The Zexe.net system – developed at the Sony Computer Science Laboratory in Paris (SONY-CSL) – consisted of a set of online applications and tools that allow small-scale communities to represent and communicate their views and daily lives on the web. Through the use of smart phones, communities in different cities around the world have published images, videos and sound recordings in Zexe.net for the last five years. Participants not only publish their daily experiences in the form of multimedia files, but they also tag them. Thus, Zexe.net proposes a novel usage for tags by letting users assign them to what we could call "slices of life". We call these web-based tools Community Memories, as they help communities represent and raise awareness about a commons (Steels and Tisselli (2008)).

The Zexe.net platform has been deployed for a range of different communities such as taxi drivers in Mexico City, gypsies in Lleida and León (Spain), prostitutes in Madrid. Two deployments of the Zexe.net system are of particular interest for TAGora: the *canal*MOTOBOY project* and *GENEVE*accessible*. In *canal*MOTOBOY*, motorcycle messengers (called motoboys) in São Paulo, Brazil, report about their journeys. In *GENEVE*accessible* handicapped people in Geneva report

about the accessibility of different locations throughout the city.

Although a first version of Zexe.net already existed before the TAGora project, the applications for the two deployments mentioned here were totally re-written in order to support folksonomies. We found that the concept and the mechanics of tagging were understood very quickly by the participants of these projects, even by those who were not technically literate. The inclusion of folksonomies in these projects greatly improved the way in which the participants dealt with emerging topics, and provided a bottom-up way for representing the issues and views of the involved groups in a much more accurate and fine-grained way. By analysing and comparing the tagging activity in *canal*MOTOBOY* and *GENEVE*accessible*, we also show how the scope of a project's focus and tag suggestion can influence the growth and diversity of folksonomies. The Zexe.net system includes the basic functionalities of folksonomies: tagging with or without suggestions, tag clouds which are viewable using different criteria (frequency or popularity), filtering searches through tags and grouping.

Within the scope of TAGora, the ideas behind the Zexe.net platform have lived on in its successor NoiseTube which extends its concepts to create a collaborative community platform to sense and tag exposure to pollution.

NoiseTube – Community Memory representing the exposure of people to noise pollution

NoiseTube is a participative sensing (Burke et al. (2006)) and tagging platform that aims to enable citizens to gather, manage, visualise and distribute data on urban noise pollution. This innovative application of collaborative tagging extends and builds upon the concepts pioneered in the Zexe.net project.

The Zexe.net project focussed on exploring new applications of tagging for the benefit of offline communities facing a variety of issues related to the sustainable exploitation of a commons. Through the use of smart phones, communities in different cities around the world have published content on the Zexe.net platform, resulting in so-called Community Memories Steels and Tisselli (2008) which help these communities represent and raise awareness about shared concerns. Starting from the 3rd year of the TAGora project we began to extend this idea by applying collaborative tagging to a new type of resource: the exposure of individual people to pollution.

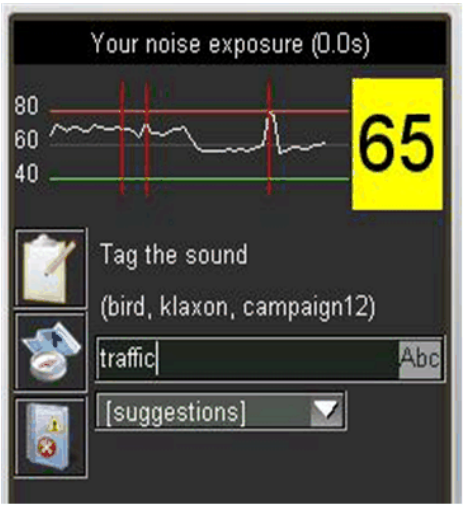
As a concrete case we studied the problem of noise pollution, resulting in the NoiseTube platform (Maisonneuve et al. (2009)). We chose the case of noise pollution because it is a major problem in urban environments, affecting human behaviour, well-being, productivity and health and because we believed it would be possible to measure noise via the microphones of mobile phones.

Platform

We extended and modified the Zexe.net platform to create a new community memory system called NoiseTube (as in "a YouTube" for noise). The principal goal was to make it possible for citizens to turn their cell phone into a personal, mobile noise sensor. The device monitors the level of environmental noise and feeds this measurement data, along with collaborative tagging data, into a centralised web-based database accessible for everyone.

The NoiseTube platform consists of an application for mobile phones and a web application. The mobile application acts an actual noise level meter which uses the microphone of the phone (or an external one) to measure exposure to noise in real time (in 1 second intervals). This information is shown on the screen of the device and is enriched with metadata such as a timestamp, GPS coordinates and manual and automatic tags, before being sent to the central web application. This web application, which is accessible through the NoiseTube website (http://www.noisetube.net), collects and aggregates the measurements from the distributed network of phone-based noise sensors and provides features to navigate, download and visualise the data in different formats.

(a) A NoiseTube contributor measuring the level of environmental noise at a construction site using her mobile phone as a personal noise sensor instrument. The data is directly sent to the NoiseTube web application to update the exposure map of the city



(b) The user interface of the mobile application. On top, the visualization of the exposure. On the right, the current L_{eq} in dB(A) with a colour representing the degree of risk (green, yellow or red). At the top left, the log the exposures with red vertical lines representing tag assignments. In the centre, a free text field to tag the exposure with a dynamic suggestion list)

Mobile application

To turn off-the-shelf smart phones into noise sensors we implemented a signal processing algorithm which computes – in real-time – accurate noise level measurements based on an audio signal. The accuracy of the measurement was evaluated with laboratory tests which were carried out in a sound-proof audio studio. In a later stage we conducted a series of real world tests in the streets of Paris, in collaboration with BruitParif[5], the official observatory of noise pollution for the Parisian region. We concluded this evaluation and calibration process with an observed average error of $\pm$ 2.5 dB(A), which we consider accurate enough for our goals.

Measuring noise exposure is not enough. We also need to identify the causes of the pollution to react on it. As people are excellent at recognising noise sources, they can annotate the measures regarding the cause or context of their exposures such as cars, aircraft or neighbours via the mobile application to inform the community about it. In fact public noise maps often provide only a very limited information regarding the source or context of noise. This sort of semantic information is collected through social collaborative tagging. This type of metadata is vital to build meaningful noise maps for both citizens and decision makers.

Using techniques to extract semantic descriptions of low level features – similar to ideas from the Ikoru project; the mobile application not only acts as a sensor but also does post-processing to detect exposure patterns and tag them automatically. For the moment two basic patterns are supported:

> **sudden high variation**: when there is a sudden high variation (+ 15 dB(A)) in a short time (< 3 seconds) the application automatically adds the tag *"sudden peak"* to the last measure.
>
> **long and risky exposure**: When the user is exposed to a high level of noise (> 80 dB(A) for a long time (> 20 seconds) the mobile application automatically adds the tag *"risky exposure"*.

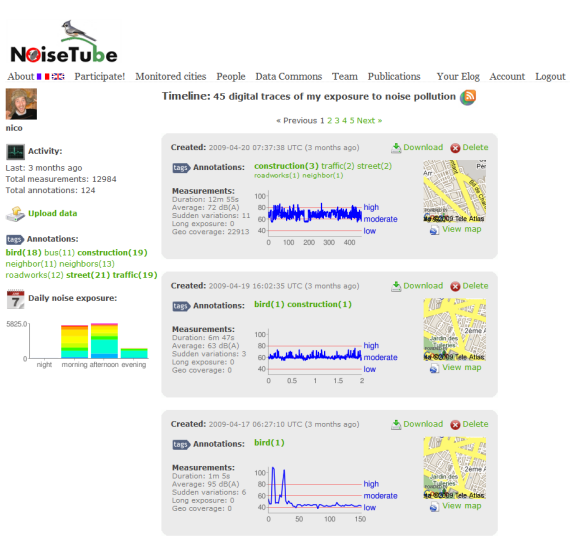
Extracting basic patterns to generate tags allows adding a semantic description of these levels of exposure and the tags can be used to power the same navigation and visualization features as the

[5] http://www.bruitparif.fr

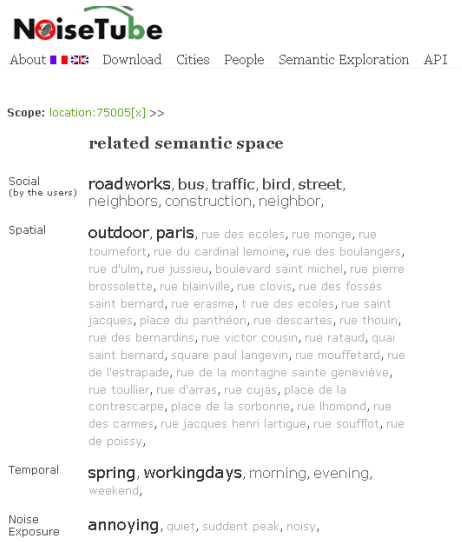
other (human) tags. In our case we can discover clusters where people had sudden high variations or long dangerous exposure without further inferences. Furthermore using the phones to distribute the computational effort allows a better scalability of the general system.

Web application

The web application provides several ways to navigate and visualise the exposure to noise of people in cities. Once the measured data are sent the server, any user can see his own contributions or exposures by going to the NoiseTube website (http://www.noisetube.net). Fig. 2.2(c) and 2.2(d) show screenshots of principal parts of the web application.



(c) The *eLog* (environmental log or exposure log) page showing a list of recorded digital traces of the exposure of a user



(d) The "'semantic exploration"' feature enables users to explore data by drilling down on a dataset by iteratively selecting tags (some automatically generated) from different semantic dimensions

Figure 2.2: NoiseTube web application

For each city an noise map can be downloaded as a KML file, which can be visualised using the Google Earth[6] application. Each map consists of multiple layers, such as the noise exposure layer, consisting of all measurements and a semantic layer with all tags. This map is constructed by aggregating all the shared contributions. The layer of tags adds a context and meaning to the physical measurements of noise pollution allowing, for instance, to identify the sources of noise.

The concept of a tag cloud is represented on a map by a pie showing the distribution of the different sources of a given area. We created a feature allowing to contextualise this tag cloud according to the geographical area displayed: if the user moves or zooms, the tag cloud is recomputed in real time.

Data in environmental pollution and exposure to it are generally not directly accessible for the public or scientists, limiting their exploitation by third parties. the NoiseTube platform provides a simple web API for publishing or accessing raw data and tags. Using this API, external users can access annotated, individual or collective noise exposure data, for example to create web mash-ups or to analyse data.

[6] http://earth.google.com

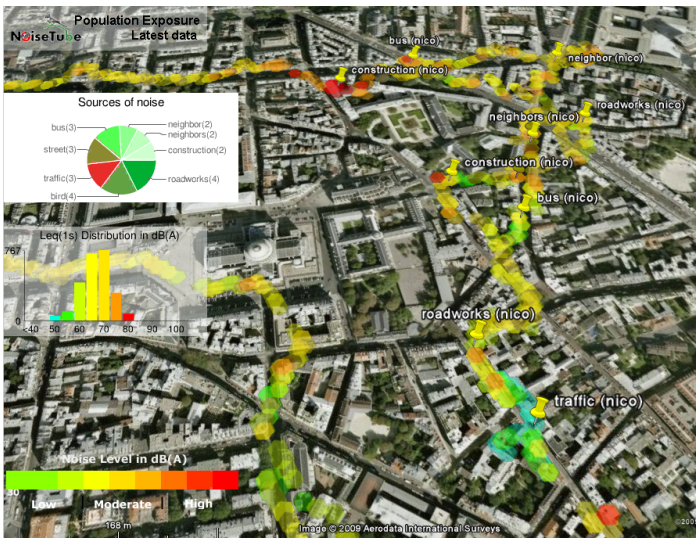


Figure 2.3: Visualisation using Google Earth representing the collective exposure to noise pollution generated by all the measurements and a tag cloud (shown as pie chart) with sources of noise

TAGnet – a tool for awareness and management of personal metadata

People are not fully aware of the metadata they use to annotate resources. Tagging requires little effort and implies no strong committment to consistency: this fosters the externalization of large bodies of metadata, and at the same time makes the structure of the metadata rather unpredictable, even for the user performing the annotation. In the context of a single user, the tag co-occurrence network exposes many of the semantic relations among tags, and between tags and the broader context defined by user's interests and experiences. Visualizing the tag co-occurrence network and allowing the user to manipulate it provides her with a sort of "semantic mirror" that can be used for awareness, for navigation, and for re-organization of metadata. To this end, an application initially not envisioned in the work-plan, TAGnet (http://www.netr.it), was designed to exploit the results of WP3 and WP4.

Features

- A Flickr user can connect to the TAGnet web site (http://www.netr.it) and insert her username. The system fetches from Flickr a list of the tags associated with each annotated photo and computes a co-occurrence network, which is subsequently visualized by using the above force-based layout engine. The user interface allows users to tune the number of tags displayed by the interface, and the threshold of co-occurrence controlling whether a link is drawn or not between two tags. The user can also dynamically exclude tags, mark two tags as "synonyms", mark a tag as a lexical variation of another tag, mark a few tags as "important", and so on. By means of these actions, supported by the user interface, users can further structure the metadata and drive the visualization towards what they think is a better representation of the categories and conceptual structures they consider relevant. The user can switch among several different layout schemes and even freeze the layout engine to arrange tags manually according to her will. The resulting (user-manipulated) visualization of the tag co-occurrence network can be uploaded to Flickr and shared by clicking on a button.

MyTag – a tool for integrating folksonomies

MyTag (http://mytag.uni-koblenz.de/) is a cross folksonomy search tool which allows for searching different resource types like photos, videos and bookmarks from different folksonomy sites in parallel. For this purpose, it queries the public APIs of different tagging systems and presents the retrieved results in a separate column for each content type (see Fig. 2.4). Besides the search across different tagging systems, it collects the search interests of registered users and offers them a personalized ranking of results. Additionally, an intelligent search assistant helps disambiguating the current search terms by grounding them to possibly relevant articles found in DBPedia.

Features

- Incorporating further platforms and content types: With Bibsonomy and Connotea, two additional platforms were introduced into the system. By incorporating Bibsonomy, MyTag now also supports the search in bibliographic references.
- Merging search result lists for the same content type: By incorporating Bibsonomy and Connotea, we now retrieve search results for bookmarks from three different platforms. This required to introduce an algorithm into Mytag that merges the bookmarks coming from Delicious, Connotea and Bibsonomy into a single result list. The technical details about the merging algorithm are available in Grabs (2009).
- Intelligent search assistant I: In a collaborative effort with the Southampton team, we introduced an intelligent search assistant into MyTag. It automatically analyzes the current search terms of the users and sends it to a disambiguation web service offered by the Southampton University. This web service grounds the search tags in articles from DBPedia and returns possible related terms. MyTag then analyzes the returned list of possible meanings of the search terms and filters those meanings which are not represented in the search results retrieved from the tagging platforms. The remaining grounded terms are then presented to the user. The user can then select the intended meaning of the search term and re-rank the current list of results so that resources corresponding to the intended meaning are ranked higher.
- Intelligent search assistant II: In Abbasi and Staab (2009) a method is proposed how to identify generalized and specialized tags. This analysis was applied on the collected Flickr data set and the resulting lists of generalized and specialized tags were also used for suggesting tags to a user during his search. But the evaluation in Scharek (2009) showed that users only seldomly found the suggested tags useful for refining their search. Because of these mixed evaluation results, this search assistant was never included into the publicly available version of MyTag.

Implementation MyTag is implemented using the Ruby on Rails framework as it supports efficient development of web-applications. The MyTag architecture realizes the model-view-controller paradigm (MVC). A view layer at the top is responsible for the interaction with the user while the control layer in the middle processes data from the model layer, e.g. by computing personalized rankings.

Two personalization features are provided for search: First, a search can be restricted to resources uploaded by the user. This feature requires that a user enters her external account names for Flickr, del.icio.us, and/or YouTube into her profile. Searching only own resources is implemented by using the corresponding feature from the integrated tagging platforms. The second personalization

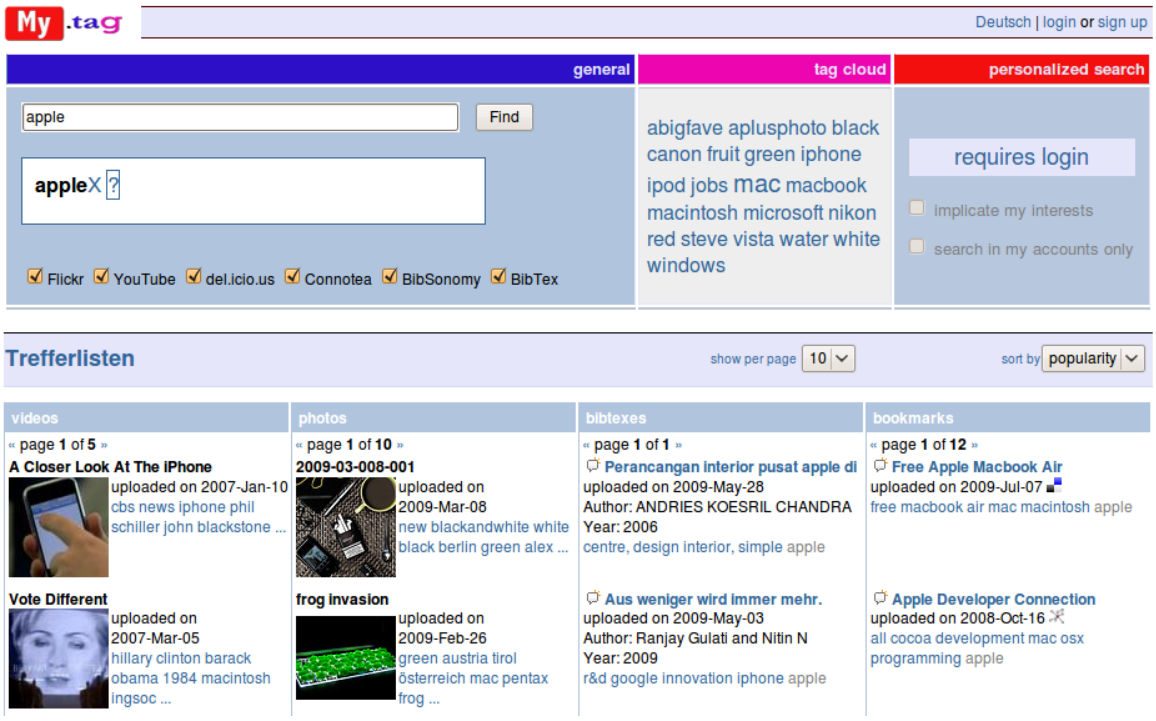


Figure 2.4: A screenshot from MyTag

feature allows for ranking search results based on the user's personomy. The personomy is automatically built based on the resources the user picks from the result set. It is modeled by a vector p of tag frequencies representing the previous search interests of the user. As it is based on the implicit feedback given by selecting from the search results, no additional user effort is required to gain personalization. Using implicit user feedback is a very promising approach to personalizing search results or web browsing in general. This feature adds an advantage compared to systems such as Flickr and del.icio.us, where personalization requires adding resources to the system, i. e. the explicit feedback of users.

Live Social Semantics

LSS was very successfully deployed at the 2009 European Semantic Web Conference (ESWC09) and HyperText Conference (HT09). It provided several services to the users that are aimed at encouraging social interactions at facilitating finding friends and people with similar interests. LSS mined real-world interactions of conference attendees using hardware and software infrastructure developed by the SocioPatterns.org project (Barrat et al. (2008)). The conference name badges of the users were equipped with active RFID badges. The RFID badges engage in radio communication, and can reliably assess the continued face-to-face proximity of two individuals. We assume continued face-to-face proximity to be a good proxy for a social interaction between individuals.

LSS provided two types of visualisations of RFID-contacts; a global visualisation, showing all users in the conference rooms and their live face-to-face contacts, and a personal visualisation, showing the accumulative view of someone's face-to-face contacts.

In previous work (Szomszor et al. (2008a)), we devised an architecture to automatically generate a list of DBpedia URIs to represent interests a person might have by reasoning over their social tagging activity. This was integrated into LSS, where any social tagging information from Delicious and Flickr is collected and converted to an RDF representation (according to the TAGora tagging

ontology[7]). This information is then used to generate a profile of interest for each user (see D4.5 for further detail).

In addition to the features above, LSS also displayed lists of online-friends that are at the conference, and recommended conference talks based on the social connection of a person, as well as on his/her community of practice.

At ESWC09, 139 of the conference attendees registered on our LSS site and together entered 246 social profiles from Delicious, Flickr, lastFM and Facebook. Out of those users, 59 entered at least one tagging account (Delicious, Flickr, or lastFM). Our policy was not to use the generated profile unless it is verified and saved by the users, to avoid publishing anything that the users might not be happy with. In the end, 31 users had a non-empty profile of interest generated for them. When generating those profiles, a total of 1210 DBPedia concepts were proposed (an average of 39 per person across the 31 profiles), out of which 247 were deleted.

When comparing the results from Delicious and Flickr, we see that 17% of concepts proposed from Delicious Tags were deleted, and 32% respectively for Flickr tags. This suggests that the accuracy of topics harvested from Delicious tags was more accurate than those from Flickr. Inspection of the concepts removed shows that Flickr was likely to suggest concepts referring to years and names.

More detail about LSS can be found in Alani et al. (2009).

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

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

2.3.1 Objectives

Examining quantitative aspects of folksonomy is a highly requested area of research. The objective of the WP was the set up of several protocols of data analysis to be performed on the raw data sets delivered by WP1. A data analysis protocol is defined by: (1) indicating a specific quantity/observable/estimator suitable of a quantitative measure on the raw data sets; (2) acquiring the existing software tools, or developing new specific tools, needed to perform the measure; (3) extracting the relevant statistical information characterizing the analyzed data sets. The aim of the data analysis was to identify and quantify emergent properties of the system in study, i.e. properties that can not be simply inferred from the behavior of the single agent. Beyond to suggest the collection of new or more refined raw data, the results of the data analysis were used to

1. identify general features common to the different systems in study;
2. characterize/discriminate the specific features of different systems in study;
3. orient the modelling phase of the research project (see WP4);
4. provide benchmarks to test/improve existing systems or to suggest the creation of new more performing systems.

Assessment and evaluation elements. A measure of success for this workpackage cannot be defined independently from that of WP4. Both workpackages act in a sort of loop implementing the standard Complex Systems Science approach. On the one hand “universal features” extracted from the experimental data are used to check the theoretical constructions. On the other hand, original theoretical predictions can be checked against the experimental data. The specific success of this workpackage can be then expressed as the ability of extracting from the experimental data specific non-trivial features which could allow for a discrimination among different theoretical schemes as well as for suggesting correct interpretation keys of the underlying phenomena.

2.3.2 Contractors Involved

UNIK

UNIK investigated the network structure of folksonomies by adapting measures for so-called “small world networks” to the particular tripartite structure of folksonomies. Additionally, a hierarchical clustering of the tag space was done by iteratively applying the k-Means clustering algorithm. The tag clusters on the bottom level were considered as intensional descriptions of UNIK’s FolkRank algorithm which was implemented in the BibSonomy system. FolkRank generates, for a given tag, beside a ranking of the resources (i.e., publications and bookmarks) also a ranking of the users that are most related to this tag. Furthermore, spam detection and tag recommendations in BibSonomy was investigated by UNIK.

UNI KO-LD

UNI KO-LD investigated characteristics of folksonomy data from delicious and flickr with the main focus on tag distribution, tag co-occurrence and use of singular/plural forms in the datasets. The tag classification system T-ORG was developed which assigns resources to categories based on the categorization of their related tags. Machine learning methods were applied for inferencing knowledge from the gathered dataset. Background knowledge was used to improve the classification of location tags to provide better navigation support for users. Based on the observation that the sparseness of tagging data hampers the quality of search results in tagging systems different

approaches have been developed for enrichting the vector space model with inferred data, e.g. based on user or tag co-occurrence networks. In the context of MyTag some effort were done in the direction of cross-folksonomy integration.

SONY-CSL

SONY-CSL did some novel work on image analysis tools to augment tag-based browsing. Visual features from the liter- ature were evaluated, combined with the KNN classification method, and made available in Ikoru through an intuitive interface. A new approach to classify images using a genetic algorithm was developed and evaluated. Also a new approach to improve automatic music classification through signal analysis and tag statistics was developed and tested. A method for automatic inference of music tags from the analysis of the audio signal was developed. Furthemore, the data of the Armin Linke exposition was analysed.

PHYS-SAPIENZA

PHYS-SAPIENZA actively participated in the data analysis by adapting already existing tools and devising new ones adapted to the structure of folksonomies. To unravel the possible cooperative behavior of users in folksonomies some of the dynamical statistical properties of folksonomies like the statistics of inter-arrival times, tag-tag correlations, and tag dictionary evolution were analysed by PHYS-SAPIENZA. Moreover, the simultaneous temporal evolution of tag streams extracted from different folksonomies was analyzed together with UNI-SOTON, looking for synchronous user activity in correspondence of known external events. 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 its similar measure defined by taking account of the co-occurrence weights, are compared with a shuffled co-occurrence network, where the semantic correlations had been destroyed.

SOTON

SOTON concentrated on cross-folksonomy network analysis. Cross referencing and linking between data in different datasets has been investigated Tools were developed for rendering tags from different clouds to make them more comparable through various tag filtering mechanisms. Various experiments were conducted for researching and understanding how distributed tag clouds that belong to the same user can be identified, and merged towards a joints profile of interest. Finally SOTON opened the access to several of the services for cross-folksonomy integration and analysis including Tag Filtering, Sense Matching, and Profile Building. These tools and services were used in MyTag as well as in LSS.

2.3.3 Work Performed

Task 3.1 Emergent metadata statistics

An extensive description of the different 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.

The stream analysis has 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 reveals non trivial correlations and has been compared to same measures performed in book texts, as well with random shuffled streams.

The emergence of topics in folksonomy evolution has been investigated by studying the statistics of bursts in the tagging activity of several folksonomies. A reasonable hypothesis would predict that topics, mimicked by tags, maybe divided in two class according to their usage: on one hand, bursty tags related to special events or other peak of interest (e.g. “worldseries2008”), and regularly used

one, related to daily activities of users (e.g. “blog”). If this was the case, the categorization of topics could be performed (also) based on the dynamics of the corresponding tags. Inspired by methods employed in the theoretical modelling of the physics of earthquakes, we have explored in detail the statistical distribution of interarrival times of tag occurrences.

Task 3.2 Network/graph analysis

Task 3.2.1 Topological properties

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

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

Task 3.2.2 Dynamical properties

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

Task 3.3 Cluster/community identification

Our work can be grouped into three parts. First 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. k-Means was chosen as initial clustering algorithm and the semantic grounding was provided.

Task 3.4 Semantic Inference

As a preparation step for semantic inferencing we first investigated what background knowledge would be useful for which tasks. We did this by focusing primarily on data filtering, data enrichment and data classification.

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

We performed a user study for evaluating proposed and simple vector space models. The results were evaluated by 18 expert users (mostly PhD students) who were well familiar with search and image search. Each user was shown a search result page.

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

Task 3.5 Cross-Folksonomy Networks

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

Task 3.6 Collaborative tagging and emergent semantics

Task 3.6a Improving Navigation for Images

As part of the process to improve the navigation within the Ikoru application for images a closer analysis of related datasets from the deployed applications was initiated. Specifically the analysis of the data collected through the "Pheno-types/Limited Forms" installation. In this installation museum visitors were engaged in tagging in a physical space, and the exhibition was a mere extension of the virtual Ikoru Interface into the real world. The tag assignment distribution within dataset was examined, followed by a closer look at the three-mode network structure.

Task 3.6b Improving Automatic Classification of Music

The objective of this task was to study the relation between tags and content-based analysis to answer following questions: 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?

Intensive and rigorous experiments were conducted 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.

The research in automatic tagging in the music domains was continued with the design and evaluation of efficient techniques for predicting automatically tag from the analysis of acoustic signals of polyphonic music. Additionally, an original study concerning the possibility to predict Hit Songs using acoustic and manual metadata was performed, and published in Pachet and Roy (2008b).

2.3.4 Final Results

Task 3.1 Emergent metadata statistics

Studying the statistics of bursts in the tagging activity of several folksonomies, we observed that the tagging activity related to popular topics follows the same dynamics of the activity related to more specialized ones, if a suitable rescaling of the time unit is performed. As a consequence, by a mere observation of the tagging dynamics it is not possible to clearly discriminate the character of tags (bursty as opposed to regularly used ones) and, hence, infer their categorization.

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

Task 3.2.1 Topological properties

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. A model successfully 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, were compared with a shuffled co-occurrence network, where the semantic correlations has been destroyed. The non shuffled network shows a reduced size of the cores, indicating a different network topology and suggesting a hierarchical organization of tags.

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

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

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

By contrast, the cosine similarity displays a clearer signature of the influence of social interaction on semantic similarity. While in the real dataset the semantic similarity decreases for larger dis-

tance between the users in the social network, in the null model there is a very weak dependence between the two quantities. Details on this analysis are currently submitted for publication Schifanella et al. (2009).

Task 3.2.2 Dynamical properties

We have measured the average semantic similarity of neighbor users as a function of time, of the chosen network and of the adopted measurement method. We have compared the data with a suitable null model, where all statistical properties of the tag stream are maintained, but the association between tags and users have been randomly reshuffled.

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

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

Task 3.3 Cluster/community identification

In Deliverable 3.2, we presented three different studies on community detection.

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

We investigated on different aspects of semantic inferencing based on background knowledge, namely 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 and richVSM (see Abbasi and Staab (2008, 2009)) were developed for overcoming this problem by recombining the existing data and using them in Latent Semantic Analysis to identify hidden concepts.

Using enriched vector space models, we show that one can find meaningful relationships between tags and then use these relationships to reduce the sparseness in folksonomies. We find the relationships between tags based on two dimensions, first the context of the tags and second the distribution of tags. We consider two types of tag contexts, the resource context (which resources are assigned a particular tag), and the social context (which users have used a particular tag). The resource context of tags helps in finding tags which are mostly used in similar kind of resources, whereas the social context finds broad relationships between tags based on the users' interests (represented by the tags they use). We also exploit two kinds of tag distributions, 1) similar tags and 2) generalized tags. We find relationships between similar tags by using the existing *cosine* similarity measure and propose a modified overlap coefficient to exploit generalization relationships between tags.

We hypothesize that the statistic description of resources that use common tags exhibits different behavior than the statistic description of resources with uncommon tags. To test this hypothesis, we split the queried tags into three sets; having 1-10 search results, 11-50 and more than 50 search results respectively and perform experiments on these sets of queries. We also propose a method *Best of Breed* (BB), which selects an appropriate enrichment model based on the number of relevant resources related to the queried tags.

The experimental results of the large scale evaluation (150 queries total evaluated on a dataset of ~27 Million resources by 18 expert users) show that the enrichment of existing data by exploiting semantic relationships among tags helps in improving the search results. We computed precision at 5, at 10, at 15, and at 20 for each of the methods and query sets and compared to the average precision without enrichting the vector space model. Particularly for the queries which have a few relevant resources in the original data, theres was a great improvement of the results. We also observed that a significant improvement in the precision at all levels was achieved with the vector space model based on semantically similar tags using social context.

Task 3.5 Cross-Folksonomy Networks

Different services for cross-folksonomy integration and analysis were developed, some of which are openly accessible. These services include:

- **Tag Filtering:** When people tag resource, be it a web page, photo, song, or video, they are free to choose any tag(s) they please. While it has been shown that this uncontrolled behaviour does result in meaningful semantic structures, the tag-clouds of particular individuals often contain misspellings, synonyms and morphologic variety. As a result, important correlations between resources and users are often lost simply because of the syntactic mismatches in the tags they have used. The Tag Filtering service can be sent a set of raw tags (e.g. an entire tag cloud), which will be processed and filtered by the service, and a *clean* set of tags will be returned back (details in Cantador et al. (2008)).

- **Sense Matching:** The TAGora Sense Repository (TSR) is a linked data enabled service endpoint that provides extensive metadata about tags and their possible senses. When

queried with a tag, the TSR will attempt to find DBpedia.org URIs and Wordnet Synsets that correspond to the possible meanings of the tag. Since many of the tags used have multiple meanings (e.g. apple may refer to the fruit or the technology company), the TSR also provides additional metadata about DBPedia senses to assist in the disambiguation process (details are in Garcia-Silva et al. (2009)).

- **Profile Builder:** With the growth of Web2.0, it is becoming increasingly common for users to maintain a presence in more than one site. For example, one could be bookmarking pages in Delicious, uploading images in Flickr, listening to music in Last.fm, arranging social events with Facebook, etc.The Profile Builder service (currently not public) generate rich *Profiles of Interests* by bringing together and consolidating multiple folksonomy identities.

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

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

Task 3.6 Collaborative tagging and emergent semantics

Task 3.6a Improving Navigation for Images

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

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

Task 3.6b Improving Automatic Classification of Music

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).

We came up with a novel scheme to improve the accuracy of supervised classification techniques, using a kind of ensemble learning approach. Moreover, an in-depth evaluation of the technique on a large database of music (about 33.000 MP3s) was performed. The results of this study have been published in Pachet and Roy (2009). Many interesting insights were obtained, notably showing counter-intuitive effects (e.g. high-level descriptors are not necessarily more difficult to predict that low-level ones). The study about predicting hit songs using acoustic and manual metadata (Pachet and Roy (2008b)) received a great deal of attention from the music information retrieval community. However, the results were shown to be globally insufficient for completely automatic predictors. They were, therefore, not implemented in the Ikoru framework.

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

2.4 Workpackage 4 (WP4) - Modeling and simulations

2.4.1 Objectives

The goal of this workpackage has been the development of models and simulating algorithms able to capture the most relevant aspects of the tagging activity on actual systems. Such theoretical tools, however, are a preliminary ingredient for technological applications. The theoretical knowledge inspires control strategies and allows to test and implement service improvements in real systems.

Folksonomies are a superposition of several linguistic and social activities mediated by the World Wide Web. Folksonomy users provide descriptions of known objects, establish social relations and are influenced by other users in their interests and tagging activity. For each of these topics, models exist in specialized literature. Nevertheless, in folksonomies the social, cognitive and linguistic aspects are so strictly interlocked, that new models are needed to capture the properties of folksonomies.

The science of complex systems has recently generated new tools and framework to tackle the study of such large-scale social systems. Complex network theory, for example, has unveiled the ubiquitous characteristics of self-organized networks, a realm to which a collaborative tagging community naturally belongs; so-called sociophysics has successfully applied physics-inspired statistical methods to analyze the properties of communities, how new friends are made, how consensus forms and opinions spread in a social network; linguistics and computer science have provided methods to study how words and texts can be grouped in categories Castellano et al. (2009).

The aim of this Work Package, therefore, has been to merge and revise all these separate tools, and possibly unify them in a single framework. This should provide models through which one can predict the behavior of tagging users: which tags they use more often, how often they follow other users' or system's suggestions, which emergent semantics they establish by their tags. In addition, within this Work Package new approaches should be devised in the study of folksonomies as a *per se* scientific objects, with characteristic emerging properties irreductible to their separate linguistic, social and informatic aspects.

Such models, in turn, help in the construction of control strategies and tools. For example, based on such knowledge one can design spam detection systems, tag recommendation algorithms, user-generated classification methods, search and ranking methods. The bibliographic references sharing system to be developed within the project represents a test-bed for such technological tools.

2.4.2 Contractors Involved

SONY-CSL

The SONY-CSL team has analyzed the data coming from the "Phenotypes / Limited Forms" installation, an art-piece and an experiment of real-world tagging based on the public assignments of titles to user-generated collections of photo of the german photographer Armin Linke. Data about the folksonomy built during the experiment have been analyzed and studied, and fed to the database of the web-based system Ikoru.

PHYS-SAPIENZA

The PHYS-SAPIENZA team has started the seminal analysis of data coming from folksonomies by the study of the vocabulary growth. The data analysis has shown that the vocabulary growth within a folksonomy exhibits a peculiar behavior, and new theoretical models have been successfully designed in order to capture such properties.

An extensive analysis of tag-to-tag relations in folksonomies has been carried out by the team.

Methods to measure the similarity between tags have been compared. Such methods have also been modified in suitable ways, so that they have been introduced in existing systems to perform many related tasks. Tag and users recommendation algorithms and the methods for the detection of emerging semantical topics developed and implemented within TAGora have heavily relied on such comparative studies. **UNI-SOTON**

The UNI-SOTON team has focused its effort on the design and the implementation of recommending systems. The recommendation task can involve all classes of objects forming a folksonomy. UNI-SOTON has started by simulating a method for movie recommendation to users exposing their interests. Besides, the study of such interests has been carried out over multiple folksonomies, and the detection of similar profiles across many folksonomy has been extensively studied.

Then, the activity of the team has focused on the application of such ideas to existing service. For example, the multimedia meta-folksonomy MyTag has been enriched by a sense recommendation system for disambiguating queries and tags, and the Live Social Semantics (LSS) experiment, where users attending a social event are are profiled according to their cross-folksonomy interests and live social contacts and accordingly recommended possible contacts and topics, are example of such practical applications.

UNIK

The UNIK team has studied the network properties of the classical tripartite networks built upon folksonomy data. In particular, such network representation has inspired a PageRank-like search and ranking method, called FolkRank since it takes into account the users' preferences and similarities to compute the relevance of resources, tags and users. The team has verified the effectiveness of different well-known similarity measures against the data issued by their in-home folksonomy, Bibsonomy.

Bibsonomy has also been the test-bed for several recommending methods, where the team has implemented all the algorithms developed within the TAGora project as a service to the users.

UNI KO-LD

The Koblenz team has developed a model, named the "epistemic" model, which reproduces all the statistical feature of real folksonomies. The model simulates the behavior of users tagging resources according to different strategies. Both the extraction of tags from users' background knowledge base and the extraction according the modified Yule-Simon model developed by the PHYS-SAPIENZA team are encoded in the model. It has been translated into a simulating software, which has been quantitatively and successfully tested against the data produced by real folksonomies. The streams of tag produced by the Epistemic Model have been made available to the public by the Koblenz team.

As regards the control strategies, the Koblenz team has developed and hosted MyTag, a cross-folksonomy platform where multi-media resources can be browsed, searched and tagged by users. Such platform has now been provided with automatic recommending tools to improve the users' experience: for example, sense recommendation has been introduced in order to disambiguate tags which may refer to different concepts.

2.4.3 Work Performed

Task 4.1 Modeling

The first step in the development of folksonomy models has consisted in establishing suitable approaches to characterize them. A folksonomy can be represented in many ways, involving all or

only some of their fundamental components - resources, tags and users. Both approaches have been used within TAGora.

Modeling tag streams

To determine the fundamental ingredients of a folksonomy model, the data analysis has focused first on tag streams, that is, the chronologically ordered sequence of tags in a folksonomy. The traditional text-analysis tools have been applied, uncovering that non-trivial properties characterize the tag stream. In particular, the number of distinct words employed, i.e. the vocabulary, as a function of the number of tag assignments has been studied, as a standard analysis method for texts.

However, other features typically found in texts, such as the famous Zipf distribution of word frequencies, have been investigated in tag stream. Such particular distribution is linked to the semantics of the language adopted; hence, this measurement helps in detecting whether the usage of tags in a folksonomy is influenced by their mutual semantic relationships.

To study the presence of collective dynamics taking place in folskonomy, and inspired by recent developments of statistical physics Bak et al. (2002), the temporal patterns of tags occurrence have been studied. If tagging took place as a fully random process driven only by the Zipf law governing most known vocabularies, the occurrence of each tag should follow a Poissonian statistics, that is, a process in which each tag occurs at each time step with a fixed probability. Studying correlations and inter-arrival times statistics helps in detecting anomalies from such simple behavior. Discrepancies indeed can be a signature of collective phenomena, due to the social nature of these communities.

The characterization of tag streams has helped in the definition of a stream generation model. In the literature, one can find examples of such model, though they mainly refer to the generation of textual stream. The PHYS-SAPIENZA team has mutuated ideas from such existing models, and modified a well-known algorithm, the Yule-Simon model, in order to capture the peculiar time correlations of tags' usage.

Modeling co-occurrence networks

A second approach for the analysis of folksonomies consists in mapping relations between similar elements on a graph in so-called co-occurrence networks. For example, in a resource-based tag co-occurrence network, tags referring to the same resource in the same post are represented by connected nodes. Such networks are supposed to encode the semantics of concepts behind tags, although it is not clear how it has to be interpreted in order to reconstruct the semantic relations. Developing models of such a network could elucidate which basic principles are behind the observed network properties.

The latter task has been accomplished by introducing a model for tag co-occurrence networks. The model assumes that an underlying semantic network exists, such as the WordNet one, where edges encodes all possible semantical relationships between concepts.

The Epistemic Model of folksonomies

A simulator of tagging activity has been developed by the UNI KO-LD team. Such "Epistemic model" is based on the model for tag stream generation introduced above, and simulates the behavior of a real tagging user. A model-based simulator has been released and allows the public to produce synthetic and realistic tag streams to study their statistical properties and adopt them as reliable benchmark for tag-related applications.

Task 4.2 Control

Do tagging users need a recommendation service?

Analyzing the emergence of common tagging behavior would show that there is an influence of the users on each other which not only affects the local behavior while tagging a single resource. Such influence leads indeed to permanent learning effects. The existence of such permanent learning effects helps to answer the question whether mechanisms have to be found to deal with the inconsistencies in the tagging behavior of the users. For example, one may try to identify and merge singular and plural forms of nouns or the different spellings of compound words.

Experiments have been performed to analyze the emergence of such patterns in the tagging behavior of users. A first goal was to discover how users deal with compound words, which lead to the proliferation of tags referring to a same compound word. A second experiment aimed to check whether a consensus emerge in the use of singular versus plural forms in tags. A third one measured the relative frequencies of noun categories following the WordNet partition in Flickr and Delicious, to check how much the behavior of users differs from a folksonomy to another. Such experiments were carried out on the Delicious and Flickr data set collected by the TAGora consortium.

Profile-based recommendation methods

A first attempt to tackle the goal of realizing a recommending algorithm has been performed by the UNI-SOTON team, which has used the data from the Netflix rental service. By categorizing movies in the Internet Movie Database, and the interests of Netflix users, the UNI-SOTON team has studied the users' profile (based on a fraction of the rental data) and simulated the recommendation of movie titles to them, checked against the remaining part of the rental data to measure the algorithm's success rate.

Similarity-based recommendation methods

A new scheme for suggestion and navigation of tags has been implemented in Bibsonomy. This scheme is based on the systematic analysis and study of several similarity measures, which led – in particular – to the identification of measures which spot in a realiable way "synonym" or "similar" tags, suitable for recommendation to the user. After a comparison of different similarity measures available in the literature, the cosine tag relatedness measure based on the vector space of the 10000 most frequently used tags has been adopted. The top 10 most similar tags are then displayed to users, in order to stimulate them the addition of related users according to a shared semantics.

Recommending senses

MyTag is cross-folksonomy search portal enabling tag-based searching across 5 popular tagging systems. Many popular tags have multiple, ambiguous meanings, especially when different folksonomies are compared. For example, the tag "apple" is often used in the Delicious bookmarking system to refer to the computer company, but in Flickr, pictures of the fruit are often tagged with apple. To disambiguate such tags, a collaboration between UNI-SOTON and UNI-KOBLENZ has resulted in a new application implemented in the MyTag portal. It suggests possible senses for a search tag and subsequently re-ranks the results according to the specified meaning.

Linking Linke

Data have been collected during several art exhibitions where the installation has been participating during 2008 and 2009. 24000 users have participated, by assigning 190000 TAS to 2400 different images. The distribution of users, tags and resources occurrences has been measured, as well as other fundamental feature of the tripartite network constructed so.

Live Social Semantics (LSS)

Under the assumption that the tags used most often by an individual correspond to the topics, places, events and people they are interested in, this experiment sought to provide a novel dimension to the social interaction by providing people with a basis to expose their interests, both professional and personal, and see those of others sharing a same venue, e.g. a conference. Central to this idea is that these profiles can be built automatically, only requiring a short verification phase from the user. Users were able to associate their various social networking site (SNS) accounts with their conference profile through the LSS site. Using a profile building algorithm, one is able to suggest to users a list of possible interests they may want to expose to other conference participants.

2.4.4 Final Results

Task 4.1 Modeling

Modeling tag streams

To capture the basic statistical properties of tag stream, a model inspired by the well established Yule-Simon one for text generation has been successfully introduced. The model simulates the generation of tag sequences where at each time step a user introduce a new tag, or an already used one. If a user is assumed to re-use words with a probability which is inversely proportional to the time elapsed since their last usage, the model reproduces the correct statistical properties of tag streams (see Cattuto et al. (2006)).

The tag stream analysis has also discovered that tags occur in "bursts", that is, short periods of high frequency followed by long periods of relatively low activity. Such behavior reminds the one of self-organized critical systems. Inspired by such resemblance, scaling methods have been employed to discover a "unified law" for the statistics of tags time evolution, showing that high- and low-frequency tags follow the same basic dynamics, if measured in terms of suitable rescaled time units (see Capocci et al. (2009b)).

Modeling co-occurrence networks

The fundamental idea underlying our approach in the modelization of tag co-occurrence networks is that a post corresponds to a random walk (RW) of the user in a "semantic space": starting from a given tag, the user adds other tags, going from one tag to another by semantic association. It is then natural to picture the semantic space as network-like, with nodes representing tags and links representing the possibility of a semantic link. Empirical evidence on the distribution of post lengths suggests to consider random walks of random lengths, distributed according to a broad law. Analytical and numerical investigations show that sub-linear power-law-like vocabulary growths are observed in the model as in the empirical data. The co-occurrence network properties (weighted degree distribution, clustering coefficient, mixing patterns) are also in very good agreement with the observed ones in real network instances. The obtained network structures are almost independent from the specific choice of the underlying graph and from the properties of the random walks, revealing a remarkable robustness of the model (see Cattuto et al. (2009)).

The Epistemic Model of folksonomies

The Epistemic model of folksonomies distinguishes between two basic ways a user may assign tags: he may imitate a previous tag assignment from the stream or he may choose a word from his vocabulary, related to the content of the resource. The model appears to be the first one to predict the correct shape of the vocabulary growth, the frequency-rank distribution in co-occurrence streams and the frequency-rank distribution in resource streams. The research suggests that the

co-occurrence statistics does not only express semantics coming from the background knowledge of users, but also their influence by the random process of imitating previous tag assignments. A simulator, based on the Epistemic Model, has been made available online to the community via the TAGora website, where additional documentation can be found (see Dellschaft and Staab (2008)).

Task 4.2 Control

Do tagging users need a recommendation service?

By inspecting how users deal with compound form tags, one can see that the variant where the words are simply concatenated (e.g “sanfrancisco”) and the variant where more than one tag is used for the compound are by far the most popular variants. Such dominant behavior, however, has set after a transient period followed by a sudden change. This clearly shows that the users in a tagging system influence each other enough so that a permanent change of their behavior can be reached which is not restricted to the tagging of a single resource. However, a more stable behavioral pattern is observed when one examines the time evolution of the usage of plural versus singular forms in nouns, contradicting a mutual influence hypothesis toward a convergence scheme.

This may happen because the problem of using different flexion forms is not as obvious as the handling of compound words. In the case of the compound words, the users are forced by the tagging system to find a way how to deal with them while the usage of different flexion forms may be unnoticed by the users. Thus, they are not searching for a solution and subsequently do not look how other users deal with the problem.

By looking at the number of tag assignments for each of the noun categories, one observes that in Flickr the tag assignments belonging to the person and location category have a higher importance than in Delicious. More interesting, the differing importance of the categories on the level of tag assignments is not reflected on the level of distinct tags (see Baldassarri et al. (2007)).

Profile-based recommendation methods

To explore the relationship between the way a user rates movies and the keywords that are assigned to movies, two prediction algorithms guessing the rating a user would give to a previously unrated movie based on tag-clouds depitcting their interests have been devised. To test the algorithms presented, a training set from the full Netflix data dump containing the ratings of 500 randomly chosen users has been extracted.

The results show that a movie recommendation system can be built purely on the keywords assigned to movie titles via collaborative tagging. By building different tag-clouds expressing a user’s degree of interest, a prediction for a previously unrated movie can be made based on the similarity of its keywords to those of the user’s rating tag-clouds (see Szomszor et al. (2007)).

Similarity-based recommendation methods

A new recommender framework allows the integration and evaluation of different (semantic or ‘non-semantic’) recommender systems into BibSonomy. These recommender systems can be either installed locally or remotely (connected and queried via http), thus allowing other research teams to integrate their recommender systems and giving a broad base for evaluation. All incoming events and informations are logged for evaluation in a SQL database. The framework is described in more detail in Deliverable 2.5.

An practical instance of a recommendation service has been developed within TAGora. The recommendation task provides a user a scalable and computationally efficient evaluation of similarity-based ranking for tags. Such service allows an improved navigation via the suggestion of related tags after a given tag-based query. A number of suitable similarity measures have been examined, in order to choose the most appropriate method. Such survey has led to the choice of measuring

similarity by the cosine similarity between the two vectors encoding the co-occurrence tag clouds for each of them (see Cattuto et al. (2008)).

Recommending senses

The sense disambiguating algorithm has been succesfully implemented in MyTag, facilitating the navigation of the multimedia content made available by the platform (see Dellschaft et al. (2009)).

Linking Linke

The occurrence distribution of tags, users and resources in Armin Linke's database and in Del.icio.us have distinct statistical features, due to the different objective conditions externally imposed to the tagging agents. When one studies the global features of the three-partite networks, the properties of the two folksonomy appear to be much more similar, showing a degree of universality in folksonomies that goes beyond very different conditions users are exposed to.

Live Social Semantics (LSS)

The deployment of LSS at ESWC2009 was the first where all components were put together and integrated with a third-party platform for social sensing. A good number of participants got to use it (see Alani et al. (2009)). Experiments based on the LSS framework have been performed at two conferences, and a growing number of public events are asking for this service to be made available to participants. Experiments like the one performed in Turin during the ACM Hypertext 2009 conference have raised the participation of a greater number of users, and the results are currently under evaluation.

2.5 Workpackage 5 (WP5) - Dissemination and exploitation

2.5.1 Objectives

The objectives were to disseminate the research results, applications and strategies generated by the TAGora project within scientific and artistic communities, and also to communicate them to a wide, general audience. The main dissemination strategies were Web-based, through the TAGora website and its tools such as a blog and of course tagging of various resources. However, more diversified communication activities have been pursued 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. Special attention had to be paid to the applications developed within TAGora because they should be a crucial means to disseminate the research.Another important objective for the dissemination of TAGora is to reach an audience beyond the communities that are already interested or familiar with tagging.
The dissemination activity, both online and offline, includes the following:

- Publication in scientific journals and other written communication media, aiming to reach both the scientific community and the general public (refer to final PDK document for a full list of these publications);
- Presentation of research results at different types of scientific and artistic events, such as conferences, talks, workshops, courses, demos and exhibitions (refer to final PDK document for a full list of these presentations);
- Organisation of training activities focusing on a hands-on approach via direct courses, workshops;
- Creation and maintenance of the TAGora web portal, and individual application websites, which inform users about news and events related to the project;
- Linkable content to establish contact with a broad online community.
- Release of resources such as datasets and tools.

2.5.2 Contractors Involved

Over the three years all contractors have actively contributed to the dissemination of the results of the TAGora research, as well as the applications, datasets and tools that have been developed. Each contractor spent specific attention to the dissemination of their own contributions (publications, applications, datasets, etc.). Furthermore some contractors carried out additional dissemination related tasks. Here in an overview of this work per contractor:

UNIK
UNIK developed the BibSonomy application. Throughout the project period they have aimed to increase their user base and provide new services and features directly based on the consortium's research. As a result the popularity of this tool has continued to rise throughout whole period.

UNI-SOTON
UNI-SOTON, together with the SocioPatterns.org project, presented the Live Social Semantics Experiment at the European Semantic Web Conference 2009 and at the ACM Hypertext conference 2009. The objective was to illustrate the possibilities of utilising various TAGora technologies for the analysis of social connectivity of conference participants.

PHYS-SAPIENZA

PHYS-SAPIENZA had the main responsibility over the creation, hosting and maintenance of the TAGora web portal. Furthermore they took the lead over the production of the white paper and the flyer to present the project to scientific audiences and the general public. The team also created a data analysis tool which is freely available on the website.

SONY-CSL

SONY-CSL has developed several applications and has actively pursued the dissemination of TAGora results in scientific, artistic and other targeted communities. The team developed Ikoru, a prototype for tagging of mixed resources which was used as part of the artistic installation of photographer Armin Linke. This piece was exhibited at several venues around the world. Secondly, the team created the Zexe.net platform and its successor NoiseTube, both of which are aimed at applying collaborative tagging to support offline communities facing sustainability related issues. As both systems rely heavily on public participation they have and will continue to introduce tagging technologies to average citizens in different counties in Europe and beyond. Thirdly, Sony CSL organised an "open house" symposium in 2006 were technologies developed within the scope of TAGora were presented. Another such event will take place in October 2009. For this event the team is currently preparing a public interactive demo of the NoiseTube platform.

UNI KO-LD

UNI KO-LD made efforts to maximise the dissemination of TAGora results by promoting their Tagster and MyTag applications as a demonstrators of tagging technologies aimed at a wide audience. Furthermore they offered students at the University Koblenz-Landau an opportunity to actively work on TAGora related topics.

2.5.3 Work Performed

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

Project presentation

During the first year efforts were made to present the TAGora project to different audiences.

One the one hand a white paper, titled "Semiotic Dynamics in Online Social Communities", was prepared for scientific audiences with contributions from all partners. The *White Paper* is mainly intended to describe target problems and grand challenges, clearly recognized by all partners, and to openly communicate them to the scientific community. This report is available on the project web site (http://www.tagora-project.eu).

One the other hand a flyer for the general public was designed (see Fig. 2.5(a) and 2.5(b)) to convey basic facts about the TAGora project in a graphically appealing and non-technical form. 2000 copies of this flyer were distributed at a variety of events.

Dissemination through publications and presentations

Throughout the three years, members of TAGora have made sure that the breakthroughs in research about tagging reached diverse audiences, thus fulfilling the dissemination objectives.

All contractors have been publishing intensively, both in conference proceedings and in in scientific and non-scientific journals. These publications help to ensure that the results of the research done in TAGora are and remain widely available. Beside these publications findings were also presented at major conferences, talks, workshops and lectures throughout the world, providing

(a) Front face (b) Internal face

Figure 2.5: TAGora project flyer

means to achieve direct contact with interested audiences and peers. Please refer to the final PDK document for a full list of these publications and presentations.

Dissemination through the Web

The main dissemination initiatives have involved the Web, although many other media were also used. A TAGora web portal was created and maintained throughout the entire period. Its contents were regularly enriched with news, publications, tagging datasets and tools to create and analyse folksonomies. 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. The domain `tagora-project.eu` was registered and reserved for the TAGora project. A dedicated server was purchased and is currently hosted by the PHYS-SAPIENZA team.

Release of applications and datasets

Applications developed within TAGora have played a very important role in the dissemination of the TAGora project not only as tools for research, but also as dissemination agents. All developed applications – BibSonomy, MyTag, Tagster, Live Social Semantics, Ikoru, Zexe.net and NoiseTube – were presented in separate sections of the TAGora web portal. To disseminate them easily among students and fellow researchers, the source code of Tagster and the Ikoru system was published under an open source license.

BibSonomy BibSonomy, the social bookmark and publication system developed by UNIK, has played a major role in making public some of the research results and technologies developed within TAGora. The project has been announced in a number of mailing-lists, and has also been featured in different news sites. Over the 3 years 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. BibSonomy has been presented at several events including the Intl. Conf. on Data Mining 2006, the Intl. Conf. on Conceptual Structures 2006, at the Research Center L3S, and at the Steering Committee Meeting of the 6FP IP Nepomuk The Social Semantic Desktop. A reference to the TAGora project in Bibsonomy can be found in the Bibsonomy's about/projects section.

Tagster Tagster, a peer-to-peer application for decentralized tagging developed by UNI KO-LD started to disseminate in the first year to a selected group of users for testing it in a real world

setting. After the second year it seemed difficult to reach a critical mass for an efficient use, so the effort spent on Tagster was thus diverted to other tasks. Nevertheless the source code has been released enabling future investigation of P2P systems that use tagging.

MyTag and Live Social Semantics MyTag, the cross folksonomy search and recommendation tool set up by UNI KO-LD achieved both of its dissemination goals. First, it was presented at the demo and poster session of WWW 2008 and at the First Future Internet Symposium. After each of these events the team was able to see an increase of user activity on the platform. Furthermore, MyTag attracted interested students resulting in two bachelor dissertations. During the third year MyTag was used by UNI-SOTON to support the Live Social Semantics (LSS) application, meant as a large dissemination window. The LSS application was successfully deployed at two major conferences and has gathered several hundreds of users.

Ikoru Ikoru, the tag-based navigation application for images and music developed at SONY-CSL, was successfully put to use in the context of an installation by photographer and artist Armin Linke. During the running time of TAGora the installation was exhibited at various venues. In 2008 it was on display at the Zentrum für Kunst und Medien (ZKM) in Karlsruhe, Germany, the Bienal de São Paulo in Brazil and the "'Selective Knowledge"' exhibition at the Institute for Contemporary Art and Thought in Athens, Greece. Now it is still on display in the Museum of Contemporary Art in Siegen, Germany. References to the TAGora project are included both in the printed books, the catalogue of the exhibitions and in the space of the exhibitions themselves.

Zexe.net and Noisetube Zexe.net platform developed by SONY-CSL, which explored new ways to use tagging in real-world situations and for the benefit of small scale real communities having sustainability related issues, was successfully deployed in cities both in Europe and South-America. In the third year SONY-CSL to consolidate and extend Zexe.net to create a new platform called NoiseTube which focuses on the case of urban noise pollution. Like Zexe.net the platforms aims to disseminate tagging concepts and technologies in a new domain (i.e. environmental monitoring) and among new audiences. The team has aimed to reach a broad audience by promoting both platforms as novel means to support communities of concerned citizens. While the Zexe.net concept was designed to support small, targeted communities, the NoiseTube platform was designed to also reach a much larger public of potential users. A reference to the TAGora project can be found on the NoiseTube website (http://www.noisetube.net).

Concerning the dataset, a section data was created in the TAGora portal describing and publishing all the datasets.

Collaboration with artistic communities

Sony CSL 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.

A collaboration was set up between SONY-CSL and artist and photographer Armin Linke. Armin Linke has always been interested in how coherent collections could be made from his vast archive of pictures, including personal ones by the viewers themselves. A collaboration was set up in the context of a tagging experiment at the Venice IUAV university. The interaction with Armin Linke and his students continue at the Hochschule fur Gestaltung in Karlsruhe, Germany where SONY-CSL team participated in several workshops organized by professors Wilfried Kuhn and Doreen Mende and their students of the curator class. The installation 'Phenotypes/Limited Forms' which is linked to the Ikoru system, 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. The exhibition is Greece yielded 16 reviews, notably one in the International Herald Tribune (09/04/08, 'Artists question objectivity') and 21 announcements in the Greek press.

Figure 2.6: Armin linke "'Phenotypes/ Limited Forms"' installation using the Ikoru system developed by SONY-CSL, 2008/09

A second collaboration with SONY-CSL was set up for the project http://www.zexe.net created by artist Antoni Abad. Zexe.net involves different misrepresented social collectives broadcasting multimedia contents directly to an unfiltered webpage from mobile phones. The latest version, canal*MOTOBOY, which involves motorcycle messengers in São Paulo, Brazil, includes tagging. This refinement to the system was added at SONY-CSL by Eugenio Tisselli as part of the research done in TAGora. The project, canal*MOTOBOY, has been widely announced through different media, and is accessible at http://www.zexe.net/SAOPAULO. In the second 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.

Furthermore The SONY-CSL Open House, a bi-annual public symposium, was organized in October 2006 for its 10th anniversary showing the work of Armin linke in the context of TAGora. The next Parisian Open House event now has been planned for October 2009 will present NoiseTube. At the event visitors will also be given the opportunity to experience the NoiseTube system first hand by making noise measurements and taggings in the street and seeing the result of their work on an interactive map afterwards.

Training activities and outreach

TAGora aims to make public its research results, so that they can be useful in training. The team published the results in a way that is accessible by non-experts. Furthermore, the "'traditional"' media (such as newspaper and radio) were also used to communicate activities and research results attained by the TAGora team. Members of the project actively sought to present the TAGora project in different media. In each case, the output is collected and listed at the TAGora website. The applications developed within the project will also be made available to the public in different training and exhibition contexts. For instance, TAGora has been featured in the Italian Press (La Stampa, Il Corriere della Sera), news bulletins such as the INFM (Istituto Nazionale per la Fisica della Materia) or the "Bollettino Universita & Ricerca", and at Radio 24 (Il Sole 24 ore). Copies of these recorded features are available at the project's website: http://www.tagora-project.eu/

Figure 2.7: `http://www.tagora-project.eu`

outreach/. (refer to final PDK document for a full list of these publications);

2.5.4 Final Results

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 dissemination strategies of the TAGora project have effectively reached an important group of scientists and the general public. 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. (refer to final PDK document for a full list of these publications);

Regarding web-based dissemination, the TAGora website `http://www.tagora-project.eu` (Fig. 2.7) is now an important point of reference for everyone interested in the study of tagging and folksonomies offering several sections e.g. data, products, blog, outreach, research etc. The "'data"' section presents the gathered datasets and several data analysis/simulation tools. Such material can be accessible by downloading them directly from the website,through offered APIs or by sending an e-mail to the owner due to licence agreement requirements. To improve the dissemination of the datasets, they were formatted in the universal RDF format. Scraping tools used to crawl data from websites like Flickr or Delicious were not published, since these tools are now out of data due to frequent changes in the publishing style and text formatting in these websites. Additionally, our legal contacts have advised us not to provide them since they violate end-user agreements. The "'products"' section containing all the applications is also available. To disseminate them among students and fellow researchers, the source code of Tagster and the Ikoru system was published under an open source license. The "'outreach"' section focuses on the articles in the press. Finally a list of papers published by the members of the consortium can be found in the section "'publications"'. For each article, the full bibliographic information is given, together with an PDF version of the paper.

The efforts to reach wider audiences and introduce them to tagging were successful, particularly in achieving the involvement of people which did not use tags as means of classification before.Dissemination through artistic installations and targeted communities (by SONY-CSL) and dissemination through popular science media (by PHYS-SAPIENZA) are interesting ways of making the project results tangible for a wide audience. The Armin Linke exhibition in the Museum of Contemporary Art in Siegen, Germany, which will remain open until 20 September, 2009.

2.6 Workpackage 6 (WP6) - Management

PHYS-SAPIENZA - WP6 Lead Contractor

2.6.1 Objectives

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

2.6.2 Contractors Involved

The Project Management was carried out by the project coordinator as well as by the Governing Board and node contractors. Node contractors and responsible: Vittorio Loreto, PHYS-SAPIENZA; Luc Steels, SONY-CSL; Steffen Staab, UNI KO-LD; Gerd Stumme, UNIK; Harith Alani, UNI-SOTON. The Governing Board is composed by 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.

2.6.3 Work Performed

The project coordinator, Vittorio Loreto, has been 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 was responsible for the political and strategical orientation of the project and for any important decision concerning the proper operation of the Consortium.

All the contractors were 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 on during the project are reported in ..... A detailed description of knowledge management, training, and dissemination activities during the whole project is reported in the Plan for using and Disseminating Knowledge (D6.4).

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:

- **Kick-off meeting** Rome, June 29/30 2006;
- **II TAGora meeting** Paris, December 13/14 2006;
- **III TAGora meeting** Koblenz, May 15/16 2007;

- **IV TAGora meeting** Kassel, October 25/26 2007;
- **V TAGora meeting** Torino, May 7/8 2008;
- **VI TAGora meeting** Bagnovignoni (Italy), November 20-21 2008;
- **VII TAGora meeting** Rome, May 7/8 2009;

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

- PHYS-SAPIENZA and UNIK organized a bilateral meeting in Kassel, October 3-5 2006, focused on WP3.
- PHYS-SAPIENZA and SONY-CSL organized a bilateral meeting in Paris, October 6th 2006, focused on WP4.
- PHYS-SAPIENZA and UNIK organized a bilateral meeting in Rome, January 31st - February 2nd 2007, focused on WP3.
- PHYS-SAPIENZA and UNI-SOTON organized a bilateral meeting in Rome, March 19-20 2007, focused on WP4.
- UNIK and UNI KO-LD organized a bilateral meeting in Kassel, April 11-12 2007, focused on WP2.
- 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.
- PHYS-SAPIENZA and UNI-SOTON organized a bilateral meeting in Torino, March 4-6 2009, focused on WP3.
- UNI-SOTON and UNI KO-LD organized a bilateral meeting in Koblenz, March 9-11 2009, focused on WP2.
- PHYS-SAPIENZA and UNIK organized a bilateral meeting in Rome, February 9-13th 2009, focused on WP2 an WP3.
- PHYS-SAPIENZA and UNIK organized a bilateral meeting in Kassel, March 30th - April 3rd 2009, focused on WP3 an WP4.
- PHYS-SAPIENZA and UNI-SOTON organized a bilateral meeting in Torino, June 27-30 2009, focused on WP3.
- PHYS-SAPIENZA and UNI KO-LD organized a bilateral meeting in Rome, July 13-14 2009, focused on WP4.
- UNIK and UNI KO-LD organized a bilateral meeting in Wurzburg, August 18-19 2009, focused on WP2-WP4.

2.6.4 Final Results

Overall the management of the project along all its duration has been smooth with the exception of the problems related to the offsetting procedure of 72.750,27 Euro done by the Commission in the second payment of the TAGora project. Though the issue is not yet solved the coordination managed to keep the level of the scientific research to the highest standard without reductions in the planned activities.

Chapter 3

Dissemination and Use

3.1 Final Plan for Using and Disseminating the Knowledge

To assure that knowledge generated within the project was disseminated to a wider scientific and non-scientific audience we conducted the following actions:

Foster publications through standard scientific and engineering communication channels. Publish results in the best scientific journals and communicate the results of the project at top conferences. Use all the possible existing communication media to touch the largest possible audience;

Encourage the partners to organize tutorials at major conferences in the different fields that are relevant to the present project. Encourage the partners to contribute to summer schools or other educational activities that touch in particular younger students. Tutorials on subjects related to the project activities will be given by the senior scientists of the network in occasion of international workshops;

Foster the organization of workshop and conferences on themes related to the project;

Disseminate results to the press at large in order to diffuse them as widely as possible;

Establishing, maintaining, updating the project web site and project mailing lists;

Foster the exhibition of demonstrators in industrial exhibitions or in other contexts where the public at large and a broad scientific/engineering audience can get exposure to the ideas of the project. These exhibitions may also represent a platform in which the results of the project are tested;

Foster the preparation of professional dissemination material;

Maximize the availability of tools and experimental platforms developed by the partners during the project to the scientific community as large;

Disseminate the results to the other projects of the Complex System Initiative and to foster synergies and possible joint activities.

The following table summarizes the dissemination actions performed during the third year of activity of the project. We refer to the Final plan for using and disseminating the knowledge (deliverable d6.4) for the complete list of dissemination actions.

Type	Number
Publications on International Journals	15
Publications on Books and Conference Proceedings	62
Talks, lectures and conference presentations	84
Conferences, Workshops and Summer Schools	38
Press, Radio and TV	48
Joint Publications	17

An addtional source of dissemination is represented by the exploitable results, defined as knowledge having a potential for industrial or commercial application in research activities or for developing, creating or marketing a product or process or for creating or providing a service. A list of these results is reported in the following table:

Exploitable Knowledge (description)	Exploitable product(s) or measure(s)	Sector(s) of application	Timetable for commercial use	Patents or other IPR protection	Owner & Other Partner(s) involved
BibSonomy	Web site (service) and Web server (software)	Public, especially researchers	Preliminary Service was online at the beginning of the project, improvements have been done according to project schedule. No commercial application planned.	–	**UNIK**
Ikoru	Web site (service) and Web server (software)	Internet Service Provider, Software Integrators, Consumer Electronics, Cultural Institutions	Version 1 of Web site online (2007), Demonstrations (2007)	Copyright on software	**SONY-CSL**
Zexe	P2P client application and MySQL database	Public, Academic and Cultural istitutions, Small communities	–	–	**SONY-CSL**
NoiseTube	Web site (service) and Web server (software)	Public, small communities , Academic and Environmental agencies	–	–	**SONY-CSL**
Tagster prototype	P2P client application	Academic institutions, small companies	–	–	**UNI KO-LD**
MyTag	Web site (service)	Public, Academic institutions, small companies	–	–	**UNI KO-LD**
Live Social Semantics	Web site (service) and application for social events	Academic institutions, medium to large companies	–	–	**UNI-SOTON, ISI Foundation, SocioPatterns.org project**
TAGnet (now Netr)	web-based system for reflexive tag exploration	Public	–	–	**PHYS-SAPIENZA, ISI Foundation**

Chapter 4

Conclusion

Our overall evaluation of the TAGora project is extremely positive. All the major objectives have been reached and in some cases the project overperformed. No major deviations had to be reported neither from a scientific point of view nor from the management perspective. Here we report the most important achievements ordered for convenience per workpackage.

WP1 Extensive data collection from selected collaborative tagging systems (Full snapshot of del.icio.us and large scale snapshot for Flickr and Last.Fm); acquisition of existing datasets from several social websites (IMDB, Netflix, Wikipedia);

WP1 Extensive data delivery through the project portal `http://www.tagora-project.eu/data/`;

WP2 Realization of several web-based applications:

> BibSonomy (`www.bibsonomy.org`) allows users to upload their bookmarks or bibliographic references and assign them arbitrary labels, denoted "tags".
>
> Ikoru (`www.ikoru.net`) is a prototypical system that unifies browsing by tags, visual and audio features.
>
> MyTag (`http://mytag.uni-koblenz.de`) aims at solving the limitations of current tagging platforms by searching different content types like photos, videos and social bookmarks from different sources in parallel.;
>
> NoiseTube (`noisetube.net`) enables a new participatory approach to monitor noise pollution by turning mobile phones into noise sensors and making intensive use of tagging.

WP3 new concepts and tools for data analysis have been explored all along the project. In particular different statistical measures of tag and resource similarity have been investigated 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, via standard two times correlators, as well as through an analysis of tag inter-arrival times. The emerging picture shows that clustering of users' activity strongly affects the temporal evolution of collaborative tagging communities.

WP3 There has been an intense activity aimed at defining and using several measures to detect and propose "related" tags. The notion of relatedness, however, has always been unclear. We investigated in a systematic fashion several aspects of similarity and relatedness in social bookmarking systems, using data from popular systems (Del.ici.ous) as well as from systems developed by the Consortium (Bibsonomy). We investigated tag relatedness by

relating graph-based notions of similarity computed on the folksonomy graph with formal representations of knowledge such as the WordNet project (for terms) or the Open Directory project (for URLs).

WP4 Introduction of several stochastic models to explain users' activity on collaborative tagging systems. The stochastic model is meant to describe the behavior of an "effective" average user in the context identified by a specific tag, and can be stated as follows: the process by which users of a collaborative tagging system associate tags to resources can be regarded as the construction of a "text", built one step at a time by adding "words" (i.e. tags) to a text initially comprised of a small number of words (Cattuto et al. (2007)). In the same spirit an epistemic dynamic model has been introduced that can be used for simulating the assignment of tags to resources 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 selec- tion of tags from his background knowledge that he thinks are suitable for describing the content of the resource (Dellschaft and Staab (2008)).

WP4 Introduction of a first stochastic modeling scheme to mimic the construction of the tag co-occurrence network (Cattuto et al. (2009)). 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. Detailed test of this Semantic Walker Model have been conducted by considering networks with very different topologies in order to test the robustness of the approach with respect to the structure of the underlying network. In addition an implementation of the Semantic Walker Model using a real Word Association graph a proxy for a latent shared semantic graph has been considered. We considered in particular the South Florida Free Association Norms database (`http://w3.usf.edu/FreeAssociation/`).

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

WP5 preparation of a White Paper on open problems and challenges for Semiotic Dynamics in Online Social Communities.

WP5 preparation of a review article on the *Statistical Physics of Social Dynamics*, published in Reviews of Modern Physics (Castellano et al. (2009)).

WP5 realization of a portal focused on collaborative social systems, accomodating 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.

WP5 Organization of the Hypertext 2009 conference, the 20th ACM conference on hypertext and hypermedia, held in Turin from June 29th to July 1st 2009.

(`http://www.ht2009.org/index.php`). Ciro Cattuto was the general co-chair of the conference, together with Giancarlo Ruffo from the University of Torino. Andreas Hotho and Vittorio Loreto were the chairs of a track devoted to *People, Resources, and Annotations*. Hypertext 2009 was an excellent showcase of TAGora achievements, in terms of both paper presentations and technology demos.

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

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

Bibliography

Rabeeh Abbasi and Steffen Staab. Introducing Triple Play for Improved Resource Retrieval in Collaborative Tagging Systems. In *Proceedings 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 and Steffen Staab. RichVSM: enRiched Vector Space Models for Folksonomies. In *HYPERTEXT 2009, Proceedings of the 20th ACM Conference on Hypertext and Hypermedia*, Turin, Italy, 2009. ACM.

Rabeeh Abbasi, 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.

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

Per Bak, Kim Christensen, Leon Danon, and Tim Scanlon. Unified scaling law for earthquakes. *Phys. Rev. Lett.*, 88(17):178501, Apr 2002. doi: 10.1103/PhysRevLett.88.178501.

A. Baldassarri, C. Cattuto, K. Dellschaft, V. Loreto, V. Servedio, and G. Stumme. Theoretical tools for modeling and analyzing collaborative social tagging systems – a stream view. *Deliverable D4.1, TAGora Project*, 2007. URL http://www.tagora-project.eu.

Alain Barrat, Ciro Cattuto, Vittoria Colizza, Jean-François Pinton, Wouter Van den Broeck, and Alessandro Vespignani. High resolution dynamical mapping of social interactions with active RFID, 2008. http://arxiv.org/abs/0811.4170.

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

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

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

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

Claudio Castellano, Santo Fortunato, and Vittorio Loreto. Statistical physics of social dynamics. *Reviews of Modern Physics*, 81(2):591, 2009. doi: 10.1103/RevModPhys.81.591. URL http://link.aps.org/abstract/RMP/v81/p591.

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 (PNAS)*, 104:1461–1464, 2007.

Ciro Cattuto, Dominik Benz, Andreas Hotho, and Gerd Stumme. Semantic Analysis of Tag Similarity Measures in Collaborative Tagging Systems, 2008. URL http://www.citebase.org/abstract?id=oai:arXiv.org:0805.2045.

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

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

Klaas Dellschaft, Olaf Görlitz, and Martin Szomszor. Sense aware searching and exploration with mytag. In *Proceedings of ISWC-09 Poster and Demo Session*, 2009.

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

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.

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

Nicolas Maisonneuve, Matthias Stevens, Maria E. Niessen, Peter Hanappe, and Luc Steels. Citizen Noise Pollution Monitoring. In Soon Ae Chun, Rodrigo Sandoval, and Priscilla Regan, editors, *Proceedings of 10th Annual International Conference on Digital Government Research: Social Networks: Making Connections between Citizens, Data and Government*, volume 390 of *ACM International Conference Proceeding Series*, pages 96–103. Digital Government Society of North America / ACM Press, May 2009. URL http://portal.acm.org/citation.cfm?id=1556176.1556198.

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

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.

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

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

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).

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

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

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

Martin Szomszor, Ciro Cattuto, Harith Alani, Kieron O'Hara, Andrea Baldassarri, Vittorio Loreto, and Vito D.P. Servedio. Folksonomies, the Semantic Web, and Movie Recommendation. In *Workshop on Bridging the Gap between Semantic Web and Web 2.0, European Semantic Web Conference (ESWC)*, Innsbruck, Austria, 2007.

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.