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 Management 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: **Draft**

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

Contents

Chapter 1

Justification of major cost items and resources

In this report we provide a justification of the major costs incurred and resources deployed by the contractors during the whole duration of the project. The way the costs are linked to the project activities must be also described.

Section 1.6 provides a tabular overview of budgeted and actual person-months.

Section 1.7 includes a tabular overview of budgeted and actual costs.

Section 1.8 describes the major deviations from cost and person-month budget (see also the deviation sections of each WP and section 3 of the Periodic Activity Report).

Section 2 includes the Final Financial Statement (FORM C) and Audit certificate of the contractors.

Section 3 includes the summary financial report of the total costs claimed by each contractor.

1.1 PHYS-SAPIENZA

1.1.1 Brief description of the work performed and the major cost items and resources employed

First Year

WP1: (Planned man-months: 5, Actual man-months: 5) We realized an early small-scale crawling of del.icio.us. On these data we performed the analysis that lead to the publication of the first modeling scheme for Social Bookmarking systems. Later on we actively took part to the main crawling of del.icio.us. and we performed and intense activity of post-processing in order to make available the data in HDF5 format. We also managed the wiki page of the TAGora web-sites to host information about all the delivered datasets. We took part to the realization of deliverable D1.1: *Data delivery from from selected folksonomy sites*.

WP3: (Planned man-months: 12, Actual man-months: 10) We actively participated to the data analysis by adapting already existing tools and devising new ones adapted to the structure of folksonomies, in particular for del.icio.us, BibSonomy, flickr, imdb and netflix. We established several collaborations with the other partners and in particular with UNIK (for del.icio.us and BibSonomy) and UNI-SOTON (for imdb and netflix). A detailed report about our activity in this WP is reported in the Project Activity Report and witnessed by the list of publications.

WP4: (Planned man-months: 20, Actual man-months: 14) We proposed the first stochastic modeling scheme for a collaborating tagging system. On studying *del.icio.us* we started our modeling effort by adopting a tag-centric view of the system, that is we investigated the evolving relationship between a given tag and the set of tags that co-occur with it. In order to model the observed frequency-rank behavior for the full range of rank values, we introduce a new version of the "rich-get-richer" Yule-Simon's stochastic model by enhancing it with a fat-tailed memory kernel. The stochastic process we introduced is meant to describe the behavior of an "effective" average user in the context identified by a specific tag and, in our view, represents a first building block upon which other models, of further complexity, can be built. The above results were have been published in the Proceeding of the National Academy of Sciences and featured in the news, both in scientific and non-technical journals and magazines.

WP5: (Planned man-months: 5, Actual man-months: 7) We coordinated the realization of the Project Presentation Report (Deliverable 5.1) to which we added a flyer to disseminate the main ideas of TAGora in a capillary way. We managed the development of a White Paper (Deliverable D5.3) on the main problems and challenges for understanding, modeling and controlling Semiotic Dynamics processes in Online Social Communities.

WP6: (Planned man-months: 6, Actual man-months: 7) As coordinator of the project, we: (a) were responsible for the day-to-day condition of the project, (b) we organized the TAgora periodic meeting, (c) we acted as the main interface between the project and the European Commission, (d) we distributed the financial contribution to the partners, (e) we coordinated the preparation of the Periodic Activity Reports.

Second Year

WP1: (Planned man-months: 5, Actual man-months 5)
Data collected last year on del.icio.us and Flickr folksonomies were tested for consistency. As a result, we had to remove few records and many others had to be retrieved again in order to recollect missing data.

In order to perform a cross analysis between folksonomies, seen as a time ordered streams of tags, and usual texts written by single authors, i.e. streams of words, we retrieved the content of

the gutenberg project ranging from year 2001 to year 2006. A cleaning procedure was needed to eliminate spurious sections of words inserted by the system mainly at the beginning and end of documents.

In view of a publication of our collected data on the web portal we had to arrange data fields in a suitable way.

WP3: (Planned man-months: 12, Actual man-months 17,08)

The characterization of the possible cooperative behavior of users in folksonomies is a demanding task. To unravel this important point we analyzed some of the dynamical statistical properties of folksonomies like the statistics of inter-arrival times, tag-tag correlations, and tag dictionary evolution. These quantities were compared with the corresponding ones extracted from english texts written by single authors.

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

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.

WP4: (Planned man-months: 20, Actual man-months 18)

We introduced and studied a stochastic model reproducing the structure of the tag co-occurrence network. The model is based on a collective exploration of a latent semantic graph and the corresponding co-occurrence network is studied and compared with real data from del.icio.us and bibsonomy data-set. Beyond the observed features, the study suggested further, refined measures (tag context similarity distributions) that have been perfomed on both the model and the real data-set. These measures corroborated the model results.

Moreover, the analysis of correlations in tag streams performed in WP3 has been followed by a theoretical scaling analysis, allowing to relate several distributions of inter-arrival times via scaling equations between exponents.

Further modeling activity has been devoted to the definition of null models to be compared both with data and stochastic models. For instance, permutation of data streams have been considered in order to appreciate the role of correlations.

We contributed to the design and to the implementation of a tag similarity measure (tag context similarity) in order to improve the navigability of bibsonomy. This implied the design of an incremental approach to update the tag co-occurrence matrix in a scalable fashion.

WP5: (Planned man-months: 5, Actual man-months 7)

Our dissemination activity concerned both academic and general public audiences. Vittorio Loreto has been the vice-chairman of the International Conference on Statistical Physics STATPHYS23, and the co-organiser of the Course on Statistical Physics of Social Dynamics during the International School on Complexity (Erice, Italy). He will also lead a session during the forthcoming European Conference on Complex Systems ECCS08 (Jerusalem, Israel). Ciro Cattuto has co-organised the Workshop on “Social Websites - Complex Dynamics and Structure” at the European Conference on Complex Systems ECCS07 (Dresden, Germany) and co-chaired the “social linking” track at the ACM Hypertext 2008 Conference (Pittsburgh, USA).

The dissemination activity has involved several appearances at the italian national radio broadcast programme on science and technology. Ciro Cattuto and Vittorio Loreto have also appeared in local and national events such as science festivals, targeting a more general audience.

As regard the project website, as planned in the project schedule (deliverable 5.4), we built an online portal, including additional functions with respect to the existing website of the project. We

created a data archive, where data referring to the online collaborative systems analysed throughout the project are now available to the public. Moreover, software simulators and analysis tools are now available for public download from the project website. Data and software tools are annotated and commented so that anyone, within and outside the community, can perform his/her own numerical analysis and modeling activity.

WP6: (Planned man-months: 6, Actual man-months 6)
As coordinator of the project, we: (a) were responsible for the day-to-day condition of the project, (b) we organized the TAgora periodic meetings, (c) we acted as the main interface between the contractors and the European Commission, (d) we distributed the financial contribution to the partners, (e) we coordinated the preparation of deliverables and Periodic Reports.

Third Year

WP1: (Planned man-months: 5, actual person months: 7.5)

In order to test our Semantic Walker Model (see WP4 below), we collected, analyzed and reformat the South Florida Free Association Norms database (freely available at http://w3.usf.edu/FreeAssociation/ website) that includes roughly 60000 associations. Though this is not precisely a folksonomy database its adoption has revealed crucial as a proxy for a latent shared semantic graph.

WP3: (Planned man-months: 6.92, actual person months: 15.5)

We studied the bursty tagging activity occurring in a folksonomy in order to understand whether different bursty events, which are the result of the apparently uncorrelated action of users, might lead to any loose tag categorization.

We examined, by using the tool of the Inverse Partecipation Ratio, whether users reach a spontaneous consensus on the description of resources in terms of significant tags.

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

In order to understand whether tag order in a post carries any semantic meaning we analyzed a directed version of the cooccurrence network of tags.

WP4: (Planned man-months: 28, actual person months: 40)

We extensively analyzed the Semantic Walker Model (SWM), mainly focusing on its robustness and dependence on the underlying network, as also suggested by the reviewers of the project. We also tested the SWM with the real network of associations available at the South Florida Free Association Norms database, described above in the WP1 paragraph.

We have studied the statistical properties of the post structure in real folksonomies. The better characterization of the post structure has been used in a further improvement of the Epistemic Model and as a reference for the SWM.

Together with UNIK, we set up the theoretical basis of the user recommendation system in Bibsonomy, by studying various tag-cloud based semantic distances.

WP5: (Planned man-months: 1, actual person months: 3)

In connection with UNI-SOTON we carried the Live Social Semantics Experiment at the European Semantic Web Conference 2009 and at HyperText Conference 2009, with the objective to illustrate possibilities of utilising various TAGora technologies for the analysis of social connectivity of conference participants.

We equipped our dissemination activity with publications on international journals, conference and workshop presentations and organization.

WP6: (Planned man-months: 3, actual person months: 3)
As coordinator of the project, we: (a) were responsible for the day-to-day condition of the project, (b) we organized the TAGora periodic meetings, (c) we acted as the main interface between the contractors and the European Commission, (d) we distributed the financial contribution to the partners, (e) we coordinated the preparation of deliverables and Periodic Reports.

Description of major cost items

First Year

- Personnel costs: 1 Full Professor (5 man months) and 1 Associate Professor (5 man-months) 83264,65 EURO, 2 Post-Docs (21 man-months, 70.993,1 EURO), Project Assistant (12 man-months, 22.783,15 EURO).
- Personal computers: (Server for data analysis, 11.935,00 EURO). A powerful server to be used for data analysis was purchased, for a total expense of 11.935,00 EURO. The server is a 1U rack-mounted AMD-based machine with 4 dual-core Opteron processors (total of CPU 8 cores), 32 Gb of main memory and very fast (though small) SCSI hard disks. The server has been configured with the Linux/Debian operating system for use by the PHYS-SAPIENZA team and TAGora partners. The large amount of main memory, which accounts for most of the expense, will allow us to cache in main memory the full snapshot of the del.icio.us and flickr datasets, and perform efficiently global operations over them. Such an amount of memory is also needed to run community-detection algorithms on reasonable subsets of the tri-partite folksonomy network as well as on the tag co-occurrence network. Calculation of depreciation of durable equipment have been computed on a 60 months base.
- Other minor costs (including: Travel expenses 15972,59 EURO, realization of the project logo 678,13 EURO, realization of the first version of a web-based application 1.456,00 EURO, other specific costs 561,09 EURO).
- Overheads: (101.656,88 EURO)
- Total eligible costs: (297.763,42 EURO)
- Requested Contribution: (160.923,66 EURO)

Second Year

- Personnel costs: 1 Full Professor (2,57 man-months, 30.049,20 EURO), 2 Associate Professors (9,51 man-months, 56.398,24 EURO), 3 Post-Docs (31 man-months, 104.923,52 EURO), 1 Project Assistant (6 man-months, 24.246,38 EURO), 1 Person for Technical Support (4 man-months, 7.626,87 EURO). Total: 223.244,21 EURO.
- Personal computers: 2.387,00 EURO (depreciation of the server for data analysis bought in the 1st year of activity)
- Other minor costs: Total 11.718,10 EURO (including: travel expenses, 9.039,08 EURO; cost of the audit certificate for the 1st period, 2.489,02 EURO; the rental of the meeting room for TAGora meeting in Torino, 190,00 EURO)
- Overheads: 128.901,20 EURO
- Total costs: 366.250,50 EURO
- Requested Contribution: 213.034,43 EURO

Third Year

The following numbers are rough estimated, and can only be confirmed once the Auditing is completed.

- Personnel costs: 1 Full Professor (4.8 man-months, 57.606,08 EURO), 2 Associate Professors (20 man-months, 116.957,60 EURO), 3 Post-Docs (44 man-months, 128.166,71 EURO).

- Personal computers: (Depreciation for the server for data analysis bought in the 1st year and a Computer bought in the third year, 5.320,27 EURO).

- Other minor costs (including: Travel expenses, 24.280,19 EURO approx.; Contribution to the organization of the Hypertext 2009 Conference in Turin, 8.681,82 EURO; Cost for the audit certificate for the II year, 2.702,70 EURO; Cost for the audit certificate of the III year, 2.500 EURO approx.)

- Overheads: (167.167,72 EURO)

- Total costs: (513.383,09 EURO)

- Requested Contribution: (276.236,40 EURO)

1.2 SONY-CSL

1.2.1 Brief description of the work performed and the major cost items and resources employed

First Year

WP1: (Planned man-months: 3, Actual man-months used: 1.99) The work includes setting up a user study with University of Venice to evaluate Ikoru and gather an initial data set. Further work includes the screen scraping of Last.fm web site to obtain tagging data as well as musical extracts. We also developed a Java application to download a test database from Flickr.com used for the image analysis. A big effort was made to negotiate access to the Sony's HiFind music database.

WP2: (Planned man-months: 7, Actual man-months used: 7.03) The work includes the development of the Ikoru platform. This covers the development of a Web server in C++ (40000 source lines of code - SLOC) that handles theSOAP queries, stores tagging information in a persistent storage (SQL database), handles security (login, validation of user input), converts uploaded media files, and provides a scripting interface to run analysis and visualization tools. To deploy the server software we developed a Web site in HTML and JavaScript (5000 SLOC). The Web site lets users register, browse, upload, and tag photos and audio. A server machine was purchased, installed, and hosted in Karlsruhe, Germany.

WP3: (Planned man-months 1, Actual man-months used: 7.95) Novel work on image analysis tools to augment tag-based browsing was completed and published. Visual features from the literature were evaluated and implemented. The features were 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. A new approach to improve automatic music classification through signal analysis and tag statistics was developed and tested.

WP4: (Planned man-months 11, Actual man-months used: 0.99) Work on the modeling of the HiFind database was begun. To exploit this large database (800 boolean tags for each of the 40000 music tracks) for creative purposes we attempted a new approach to model this information using the Mean Field Theory.

WP5: (Planned man-months 6, Actual man-months used: 6.04) The work includes the preparation of the Intensive Science exhibition and the Summer School at Erice, Italy. We participated in the organisation of the Collaborative Knowledge Management Workshop at the 4th Conf. on Professional Knowledge Management, Potsdam, 2007. We gave classes at the University of Venice and at the Hochschule fur Gestaltung, Karlsruhe, Germany, on tagging. We visited many Sony divisions (see details below) to introduce collaborative tagging and demonstrate Ikoru.

Second Year

WP1: (Planned man-months: 3, Actual man-months used: 2,91) The work includes the data gathering in the Zexe.net project (*canal*MOTOBOY* and *GENEVE*accessible*) and the *Phenotypes/Limited Forms* project.

WP2: (Planned man-months: 7, Actual man-months used: 6,19) Most of the time in this workpackage was used to evolve the Zexe.net platform. To adapt it for the *GENEVE*accessible* project better support for geographical location (GPS) and a new Java tagging application on the cell phones was needed. The *Phenotypes/Limited Forms* project required the development of an adapted user interface.

WP3: (Planned man-months: 11, Actual man-months used:9,65) The time was spent on the study on the automatic inference of music tags from the analysis of the audio signal.

WP4: (Planned man-months: 1, Actual man-months used: 0,66) The work includes the study of the data gathered in the Zexe.net project. We also used time to study a particular case of "strategic tagging" that happened as a protest on Flickr last summer.

WP5: (Planned man-months: 6, Actual man-months used: 6,58) Both the *GENEVE*accessible* and the *Phenotypes/Limited Forms* projects involved a lot of work with user groups, students, and general audience to present the ideas of collaborative tagging and "community memory". Further dissemination activities consisted of publications.

Third Year

WP1: (Planned man-months: 4.26, Actual man-months used: 2,5) The work includes the continuation of the gathering of data of the Phenotypes/Limited Forms project and the start of the data gathering for the project Noisetube about tagging noise pollution.

WP2: (Planned man-months: 9.6, Actual man-months used: 10,3) Most of the time was used to consolidate and extend the Zexe.net platform into Noisetube.net. To adapt it for noise pollution and the urban context, we extended the mobile application with a real-time signal processing algorithm to measure noise level and evaluated its accuracy. To aggregate tags by urban element (e.g. street, district) we improved the localization component by developing algorithm to correct the localization of tags using urban geographical data. Finally we created a new visualization using Google earth with new features. Finally to ensure the reusability of Ikoru, we developed a lightweight API for Flickr images.

WP3: (Planned man-months: 12.07 Actual man-months used: 7.1) The time was mostly spent on 1) atuomatic tag prediction from acoustic signals, and 2) so-called Description-Bsaed Design. We also analysed the data of the Armin Linke exposition.

WP5: (Planned man-months: 6.96, Actual man-months used: 8.70) Dissemination activities consisted mainly in publications: 3 papers for Noisetube, 4 papers for music-related works. The Phenotypes/Limited Forms expositions in different places involved also a lot of works.

Description of major cost items

First Year

- Personnel costs: Professor and Researchers (13 man-months, 97.447,27 EURO), Associate Researchers (11 man-months, 50.711,53 EURO), Total 148.158,80 EURO

- Personal computers: 1.228,68 EURO

- Other minor costs 11.095,08 EURO (including: travels 4.714,44 EURO, consumables 6.380,64 EURO)

- Overheads: 85.240,62 EURO

- Total costs: 245.723,18 EURO

- Requested Contribution: 110.487,67 EURO

Second Year

- Personnel costs: Professor and Researchers (13,44 man-months, 126.576 EURO); Associate Researchers (12,55 man-months, 55.973,76 EURO); Total 182.549,76 EURO

- Personal computers: 2.994,70 EURO

- Other minor costs 9.515,29 euro (including : travels 4.792,72 EURO, consummables 4.722,57 EURO)
- Overheads: 162.065,00 EURO
- Audit costs: 1.500,00 EURO
- Total costs: 358.624,74 EURO
- Requested Contribution: 156.185,36 EURO

Third Year

The following numbers are rough estimated, and can only be confirmed once the Auditing is completed.
- Personnel costs:

- Senior Researcher and Researchers: 8,31 man-months; 69.595,25 EURO (estimated)
- Associate Researchers: 17,50 man-months; 80.438,52 EURO (estimated)
- PhD Student: 2,78 man-months, 3.880,13 EURO (estimated)
- Total: 153.913,90 EURO (estimated)

- Personal computers: 2.117,09 EURO (estimated)
- Travel costs: 2.599,05 EURO (estimated)
- Overheads: 136.195,65 EURO (estimated)
- Audit costs: 1.500 EURO (estimated)
- Total costs: 296.325,68 EURO (estimated)
- Requested Contribution: 126.326,97 EURO

1.3 UNI KO-LD

1.3.1 Brief description of the work performed and the major cost items and resources employed

First Year

WP1: (Planned person months: 1; Actual man-months used: 1.5) Koblenz participated with several machines in the distributed del.icio.us crawl led by Kassel. Koblenz also developed a distributed and platform-independent flickr crawler. It was extensively tested to ensure that a consistent dataset will be retrieved. Different suggestions from partners were integrated to further improve the crawler and its crawling strategy. Furthermore, the necessary infrastructure was set up (i.e. a server with Postgress database) and supervised during the crawling activity.

WP2: (Planned person months: 5; Actual man-months used: 7.6) A prototype of *Tagster*, i.e. the peer-to-peer folksonomy system, has been developed. It provides basic tagging functionalities for personal data and networking support to exchange the tagging metadata in the network with other peer. Tagster has been implemented in Java to run it on virtually any platform. To support distributed tagging data statistics we integrated *Bamboo*, an open source distributed hashtable implementation. Additionally, a client interface has been implemented that provides a file browser-like interface to tag, share and browse the data.

WP3: (Planned person months: 5; Actual man-months used: 4.6) We investigated characteristics of folksonomy data from delicious and flickr collected during the crawling activities. The main focus was on tag distribution, tag co-occurrence and use of singular/plural forms in the datasets. The tag classification system T-ORG has been developed which assigns resources to categories based on the categorization of their related tags. T-ORG uses ontologies and the Google API in conjunction with pattern matching to find categories for tags.

WP4: (Planned person months: 4; Actual man-months used: 1) Contributions to this work package include some initial work on semantics extractable from the folksonomy datasets. We have been investigating the use of compound words and certain flexion forms of words over time. Moreover, we have been analyzing the vocabulary usage and richness in delicious and flickr datasets in respect to specific noun categories as found in Wordnet.

WP5: (Planned person months: 1; Actual man-months used: 0.8) Dissemination activities generally consisted of publications about the ongoing research efforts. Additionally, our work in TAGora has also been presented in workshop talks.

Second Year

WP1: (Planned person months: 1, actual person months: 1) Koblenz extended the Flickr dataset by additionally crawling photo comments and users information about membership in groups, their favorites and contacts. More data was gathered from Tagster, i.e. tagging meta data of the users and query logs, and also query logs from MyTag.

WP2: (Planned person months: 5, actual person months: 4,4) Tagster was extended and improved with respect to the user interface and the management of distributed statistics. The latter one includes extensive experiments comparing the performance of the new PINTS algorithm with basic statistic update strategies.

WP3: (Planned person months: 5, actual person months: 7,8) Contributions to this work package include the application of machine learning methods for inferencing knowledge from the gathered dataset. Moreover, background knowledge was used to improve the classification of location tags to provide better data navigation support for users. Finally, the development of MyTag is also contribution to the work on cross-folksonomy networks.

WP4: (Planned person months: 4, actual person months: 5,5) We investigated the basic properties of user behavior in tagging systems with respect to imitation and use of background knowledge. The resultig tagging model simulates the adding of tags to a folksonomy. It was evaluated in terms of capturing the typical characteristics of tagging data.

WP5: (Planned person months: 1, actual person months: 1,2) Dissemination activities generally consisted of publications about the ongoing research efforts. Additionally, our work in Tagora has also been presented in workshop talks.

Third Year

WP1: (Planned person months: 1, actual person months: 1.0)

After the major folksonomy crawling activity was finished in the second year, we collected some data from our MyTag system, i.e. query logs and the generated personomy data of the users, in order to see whether we can find similar properties. Additionally, in order to make the data sets available to the public, we checked them for consistency and did some format transformations where necessary.

WP2: (Planned person months: 5, actual person months: 5.7)

We continued the work on the MyTag system by improving the result ranking mechanism and the general implementation to achieve better performance and usability. In cooperation with Southampton, we integrated a search assistant in MyTag which provides search term disambiguation based on concepts extracted from wikipedia. The work on Tagster was discontinued as suggested by the reviewers.

WP3: (Planned person months: 5, actual person months: 6.5)

Based on the observation that the sparseness of tagging data is hampering the search result quality in tagging system we have been investigated ways to overcome this problem by enriching the vector space model with inferred data, e.g. based on user or tag co-occurrence. The results were evaluated with respect to precision and recall.

WP4: (Planned person months: 4, actual person months: 10.2)

The epistemic dynamic model was further extended to integrate the simulation of complete tag postings. A more fine-grained evaluation was done to measure the difference between simulated an real tag frequency in the tag streams. Furthermore, we investigated the classification of landmark photos in the Flickr dataset. An extensive evaluation was done with a large photo dataset.

WP5: (Planned person months: 1, actual person months: 2.0)

The dissemination activity was mainly concentrating on publications at different international conferences and presentations of the MyTag system to attract more users.

Description of Major Cost Items

- Costs funded by UNI KO-LD: (46 person months, 302.400 EURO)

First Year

- Personnel costs: PhD Student and Junior Staff (15.5 person months, 48.641,41 EURO)
- Personal computers: (857,45 EURO)
- Other minor costs (including: travel expenses, 6.004,86 EURO)
- Overheads: (11.100,74 EURO)
- Total costs: (66.604,46 EURO)
- Requested Contribution: (66.604,46 EURO)

- Permanent Staff costs: Professor (1 person months, $\sim$ 8.000 EURO)

Second Year

- Personnel costs: PhD Student and Junior Staff (19,9 person months, 76.905,00 EURO)
- Personal computers: 1.305,54 EURO
- Other minor costs 9.633,32 EURO (including: travel expenses)
- Overheads: 17.431,77 EURO
- Total costs: 105.275,63 EURO
- Requested Contribution: 105.275,63 EURO

- Permanent Staff costs: Professor (1 person months, $\sim$ 8.000 EURO)

Third Year

The following numbers are rough estimated, and can only be confirmed once the Auditing is completed.
- Personnel costs: PhD Student and Junior Staff (25.4 person months) 96.973,82 EURO
- Personal computers: 1.569,59 EURO (depreciation)
- Other minor costs $\sim$16.300 EURO (including: travel and consumables)
- Overheads: $\sim$25.000 EURO
- Total costs: $\sim$140.000 EURO
- Requested Contribution: $\sim$140.000 EURO

- Permanent Staff costs: Professor (0 person months)

1.4 UNIK

1.4.1 Brief description of the work performed and the major cost items and resources employed

First Year

WP1: (Planned person months 1; Actual man-months used: 2.5)

Lead of the del.icio.us crawl. Social bookmarking data have been crawled from del.icio.us and are available only for use within the TAGora project. The del.icio.us data have been crawled from November 10 till 24, 2006. The crawling was supported by all participants of the TAGora project and coordinated by the University of Kassel. The crawl was coordinated by a central server in Kassel. We crawled completely the corresponding user pages (including all follow up pages when a user page surpassed 5000 entries).

Contribution of BibSonomy benchmark datasets. 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 Miranda Grahl (mgr@cs.uni-kassel.de) for access to our server on https://www.kde.cs.uni-kassel.de/bibsonomy/dumps/2006-12-31.tar.gz . The dataset includes data from approximately 400 users, 12.000(different)/140.000(all) tags and 39.000 resources and can easily be loaded into a mysql data base.

WP2: (Planned person months 5; Actual man-months used: 4)

Work on improved version of BibSonomy. The following extended functionalities have been added to BibSonomy: relations between tags, keyboard shortcut for BibSonomy posting in firefox, import and export of tags, OpenURL support, tag editor, OWL output, tag hierarchy, gnome desktop integration, scrapers for ACM Digital Library and CiteSeer, spam filter, logging of copy button, improvement of basket and group functionalities, tutorials, faq, extended help pages, migration to a new server to increase hardware redundancy, password forgotten functionality, improved relation management, information extraction for publications in unstructured text, customizable export formats (including CSV for spreadsheets, HTML, RTF for Word and other text processors, DocBook XML), fulltext search.

WP3: (Planned person months 5; Actual man-months used: 3) (joint work with PHYS-SAPIENZA)

Analysis of the network structure of folksonomies. We have investigated the network structure of folksonomies. To that end, we adapted measures for so-called “small world networks” which have been used on a wide variety of graphs in recent years, to the particular tripartite structure of folksonomies and show that folksonomies do indeed exhibit a small world structure. Two large scale folksonomy datasets have been analyzed with this approach.

WP4: (Planned person months 3; Actual man-months used: 1)

Analysis of topic-specific trends in folksonomies over time. We have analyzed the emergence of common semantics by exploring topic-specific trends in the folksonomy. Our approach is based on our *FolkRank* algorithm. Compared to pure co-occurrence counting, FolkRank takes also into account elements that are related to the focus of interest with respect to the underlying graph/folksonomy. In particular, FolkRank ranks synonyms higher, which usually do not occur in the same bookmark posting together. We have described a general ranking scheme for folksonomy data. The scheme allows in particular for topic-specific ranking. We introduced a trend detection measure which allows to determine which tags, users, or resources have been gaining or losing in popularity in a given time interval. Again, this measure allows to focus on specific topics.As the

ranking is solely based on the graph structure of the folksonomy – which is resource-independent – we can also apply it to any kind of resources, including in particular multimedia objects, but also office documents which typically do not have a hyperlink structure per se. It can even be applied to an arbitrary mixture of these content types. Actually, the content of the tagged resources will not have to be accessible in order to manage them in a folksonomy system. Finally, we have applied our method to a large-scale dataset from an actual folksonomy system.

WP5: (Planned person months 2; Actual man-months used: 1)

Promotion of BibSonomy. BibSonomy has been presented at conferences workshops and on mailing lists, including dbworld, kdnet-members@iais.fraunhofer.de, wi@aifb.uni-karlsruhe.de, ak-kd-list@aifb.uni-karlsruhe.de, fgml@cs.uni-kassel.de, fg-db@informatik.uni-rostock.de, fca-list@aifb.uni-karlsruhe.de, orgmem@aifb.uni-karlsruhe.de, dl@dl.kr.org, kaw@science.uva.nl, community@mlnet.org, web_graph_algs@yahoogroups.com, webir@yahoogroups.com, ontoweb-list@lists.deri.org, semanticweb@yahoogroups.com, seweb-list@lists.deri.org, cg@conceptualgraphs.org, kweb-all@lists.deri.org, all-prolearn@agws.dit.upm.es, ml@isle.org, AI-SGES@JISCMAIL.AC.UK, machine-learning@yahoogroups.com, mlearn@googlegroups.com, Web-Mining@googlegroups.com, Machine-Learning@googlegroups.com, Data-Mining@googlegroups.com, INDUCTIVE@LISTSERV.UNB.CA.

Presentation of research results at conferences. We have presented our results at several international conferences and workshops.

Second Year

WP1: (Planned man-months: 1, actual person months: 1) Preparation of the quarterly BibSonomy dumps, and providing (password protected) access on the BibSonomy web server. Management of requests for dumps.

WP2: (Planned man-months: 5, actual person months: 4)

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

WP3: (Planned man-months: 5, actual person months: 5)

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

Within BibSonomy, we have implemented the FolkRank algorithm, which 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. This results in a community of interest around every tag in BibSonomy, together with a value of participation. The community is displayed on the web page, and is returned in the BuRST format. Furthermore, we are currently shifting our focus on spam detection and tag recommendations in BibSonomy. This is also part of our dissemination activities,

see the ECML/PKDD 2008 Discovery Challenge at http://www.kde.cs.uni-kassel.de/ws/rsdc08/. The most promising approaches will be implemented in BibSonomy in the third year of the project.

WP4: (Planned man-months: 3, actual person months: 8)

In a study, we have compared search in social bookmarking systems with traditional web search. In the first part, we compared the user activity and behaviour in both kinds of systems, as well as the overlap of the underlying sets of URLs. Our experiments are performed on data of the social bookmarking system Del.icio.us and on rankings and log data from Google, MSN, and AOL. In the second part of the study, we have transformed search engine logs into the structure of folksonomies. The structure of the resulting 'logsonomy' can then directly be compared to a folksonomy of a social bookmarking system. In particular, we have compared logsonomies from MSN and AOL search logs with a snapshot of the folksonomy of the bookmarking system Del.icio.us. For details, see Deliverable 4.2.

In a second track, we have transferred supervised learning to a social bookmarking setting to identify spammers. We have presented features considering the topological, semantic and profile-based information which people make public when using the system. The dataset used is a snapshot of the social bookmarking system BibSonomy and was built over the course of several months when cleaning the system from spam. Based on our features, we have learned a large set of different classification models and compare their performance. For details, see Deliverable 4.3.

In a joint work with La Sapienza, we also studied different tag similarity measures, details see there.

WP5: (Planned man-months: 2, actual person months: 4) organisation of travel to conferences, presentation of research results at national and international conferences, preparation of the (ongoing) Discovery Challenge at ECML/PKDD 2008.

Third Year

WP1: (Planned man-months: 1, actual person months: 1.5)
We continued publishing BibSonomy dumps every half year. For the ECML PKDD Discovery Challenges 2008 and 2009, we generated additional test and training data sets.

WP2: (Planned man-months: 5, actual person months: 6)
The functionality of BibSonomy was further enhanced, as described in Deliverable 2.5. In particular, we implemented a logging framework and a spam framework, and set up the design of the recommender framework. The latter had to be implemented outside of Tagora, financed by a national follow-up project, due to the limited Tagora resources. The frameworks were the technical platform for the experiments performed in WP 4.

WP3: (Planned man-months: 5, actual person months: 0)
Following the recommendations of the reviewers, the resources were shifted to WP 4.

WP4: (Planned man-months: 3, actual person months: 7)
Our work in WP4 covered the following topics: measures for the semantic similarity of tags, tag recommendations, the analysis of user behavior, and spam detection.

WP5: (Planned man-months: 2, actual person months: 2.25)
The BibSonomy data were the basis for the ECML PKDD Discovery Challenges 2008 and 2009, as well as for the Viszards session of the Sunbelt 2009 conference. BibSonomy has thus reached a broad recognition in the research community.

Description of Major Cost Items

First Year.

- Costs funded by UNIK: (45 man-months, 302.400 EURO)

- Personnel costs: PhD Student (11.5 man-months, 50.079,98 EURO).

- Personal computers, depreciation for first project year only: 3.857,89 EURO

- Other minor costs 6.247,62 EURO

- Overheads: 12.037,10 EURO

- Total costs: (72.222,59 EURO)

- Requested Contribution: (72.222,59 EURO)

- Permanent Staff costs: Professor (1.5 man-months, $\sim$ 8.250 EURO), 1 Post-Doc (1.5 man-months, $\sim$ 6.000 EURO)

Second Year.

- Personnel costs:

PhD Students (20 man-months), Technician (0.5 man-months), Junior Staff (1.5 man-months). 79.736,71 EURO

- Personal computers: 4.040,07 EURO (depreciation)

- Other minor costs 2.462,93 EURO (including: Travel + Consumables)

- Overheads: 17.247,94 EURO

- Total costs: 104.776,55 EURO

- Requested Contribution: 104.776,56 EURO

- Permanent Staff costs: Professor (1.5 man-months, $\sim$ 8.250 EURO), 1 Post-Doc (1.5 man-months, $\sim$ 6.000 EURO)

Third Year.

The following numbers are rough estimated, and can only be confirmed once the Auditing is completed.

- Personnel costs: PhD Students (16,75 man-months), Technician (7,5 man-months), Junior Staff (11,75 man-months), $\approx$ 108.250 EURO).

- Personal computers: ($\approx$ 2600 EURO depreciation)

- Other minor costs (including: travel expenses, consumables, $\approx$ 17.500 EURO; etc.)

- Overheads: ($\approx$ 25.700 EURO)

- Total costs: ($\approx$ 154.050 EURO)

- Requested Contribution: ($\approx$ 125.400 EURO)

- Permanent Staff costs: Professor (1 man-months, $\approx$ 9.000 EURO)

1.5 UNI-SOTON

1.5.1 Brief description of the work performed and the major cost items and resources employed

First Year

WP1: (Planned man-months: 6; actual man-months: 5) Crawled music album charts from Top40-charts and participated in the del.icio.us crawl with two machines, and with the Flickr crawl with one machine. Also obtained date from Netflix and IMDB. All the data is ontologically represented and accessible via SPARQL queries. We build a dedicated ontology for each dataset and bridged between some of the entities across various ontologies with semantic relations.

WP3: (Planned man-months: 1; actual man-months: 0) No official work has been done yet on this package. However, some cross referencing and linking between data in different datasets has been done to allow us to run some of the analysis described in Task 4.2.4.

WP4: (Planned man-months: 4; actual man-months: 4) Reviewed a wide range of existing recommendation strategies and systems and reported our findings in D4.4. We experimented with recommending movie predictions for Netflix, by integrating with information from IMDB describing films and their sets of tags. Initial results showed good potential for this type of integration where data from external folksonomies can be used to enhance recommendations.

WP5: (Planned man-months: 1; actual man-months: 1) A paper was published at an ESWC workshop to detail our initial experiments on semantic recommendations and current results. We also co-organized a workshop at the World Wide Web conference in 2007 to investigate the role and suitability of certain Web 2.0 features (e.g., tagging, community interaction) for constructing knowledge

Second Year

WP1: (Planned man-months: 6, actual man-months: 6) A highly skilled final year computer science student was hired to help gather data for two months over the summer of 2007. The student helped with writing scripts to collect additional data from various resources, such as Last.fm, UK top charts, Google News, flickr, and del.icio.us. We have also spent a good share of time on collecting information on cross-folksonomy accounts and on generating RDF networks for this data.

WP3: (Planned man-months: 2, actual man-months: 4) Most of our research this year was focussed on issues related to cross-folksonomy networks (Task 3.5). We developed tools for rendering tags from different clouds to be make them more comparable through various tag filtering mechanisms. We have also ran various experiments aimed at researching and understanding how distributed tag clouds that belong to the same user can be identified, and merged towards a joints profile of interest (detailed in D3.5).

WP4: (Planned man-months: 4, actual man-months: 5) We have spent much time and effort investigating more closely what the qualitative and quantitative gains of creating cross-folksonomy user-interest profiles from a recommendation point of view. Some of the results of the WP3 work described above was used to experiment with news recommendations, which verified the benefits of (a) generating profiles of interest from cross-folksonomy analysis, and (b) using these profiles in recommendation systems.

WP5: (Planned man-months: 2, actual man-months: 3) In this second year of the project, we published 1 workshop paper, 1 conference paper, 1 journal paper. Another paper has been submitted to a conference and awaiting notification. The workshop paper was published in collaboration with

a team from the Universidad Autónoma de Madrid. The conference paper was published at the Hypertext conference of 2008, and covered the work carried out jointly by UNI-SOTON and PHYS-SAPIENZA. We also published an article in IEEE Intelligent Systems which was coauthored with the Stanford Medical Informatics group.

Third Year

WP1: (Planned man-months: 4, actual man-months: 4) Effort on data collection was focused in the last year of the project on supporting our work on cross-folksonomy integration and analysis, and on feeding into the Live Social Semantics application (D4.5). Data We have also generated much data by collecting and processing Wikipedia pages to support our Tag Sense recommendation service. Additionally, data was gathered on-the-fly by our several open services for generating profiles of interests.

WP2: (Planned man-months: 0, actual man-months: 1.5) Building the Live Social Semantics application was not planned for in the proposal. However, designing and deploying the services that LSS is built one is core to our work in WP3 and WP4. Since we were asked to chair the Semantic Technologies track of ESWC 2009, it was only natural to put our services together to realise LSS, and to use it as a great dissemination opportunity for TAGora technologies. Approximately one and a half months in total were spent on architecting and implementing the back end of LSS which integrates our many services, and on building the LSS front end website.

WP3: (Planned man-months: 1, actual man-months: 2.5) We have focused our efforts this year on developing and opening access to several services for cross-folksonomy integration and analysis. These services include Tag Filtering, Sense Matching, and Profile Building. We have also extended our tool for generating profiles of interests. These tools and services were used in MyTag as well as in LSS. Please refer to D4.5 for further information and detail.

WP4: (Planned man-months: 4, actual man-months: 4) Several recommendation services were developed and deployed; for recommending interests, tag senses, and conference talks. These services builds on the one mentioned in WP3 above to provide various recommendations, and were deployed and tested in MyTag and LSS.

WP5: (Planned man-months: 1, actual man-months: 2) In the third and last year of the project, we have published 1 workshop paper, and 1 conference paper. We have coorganised one workshop at HT2009 with all other TAGora partners. We are also coorganising a workshop with SMI, Stanford University, on collaborative knowledge that is accepted at ISWC 2009. The workshop paper was coauthored with the Facultad de Informaticá, Universidad Politécnica de Madrid. Additionally, we have launched a major dissemination event at ESWC 2009 and HT 2009 where we deployed the Live Social Semantics experiment.

Description of Major Cost Items

- Costs funded by UNI-SOTON: (36 man-months, 252.840 EURO)

First Year

Purchased a dual processor, 32 GB RAM, and 2 TB hard disk server to store the data, for $\sim$ 6.000 EURO. Cost of salaries for this year was $\sim$35.000 EURO (9 man-months).

- Personnel costs: Post Doc (9 man-months, 35.257,22 EURO)
- Personal computers: (13.019,34 EURO)
- Other minor costs (including: travel expenses, 9.863,11 EURO)
- Overheads: (11.627,94 EURO)

- Total costs: (69.767,62 EURO)

- Requested Contribution: (69.767,62 EURO)

- Permanent Staff costs: 1 Professor (0.5 man-months, ∼7.000 EURO), 2 Senior Research Fellows (3.5 man-months, ∼45.000 EURO)

Second Year

- Personnel costs: Post Doc (12 man-months, 41.825,65 EURO), Senior Researcher (5 man-months, 22.060,87 EURO), Research Assistant (2 man-months, 3.553,67 EURO).

- Personal computers: 75,74 EURO

- Other minor costs 6.423,55 EURO (including: travel and consumables,

- Overheads: 14.787,89 EURO

- Total costs: 88.727,37 EURO

- Requested Contribution: 74.070,06 EURO

- Permanent Staff costs: Professor (0.5 man-months, 7.000 EURO), 2 Senior Research Fellow (3.5 man-months, 45.000 EURO).

Third Year

The following numbers are rough estimated, and can only be confirmed once the Auditing is completed.

- Personnel costs: Post Doc (12 man-months, 36.917,64 EURO), Senior Researcher (2 man-months, 8.588,18 EURO).

- Personal computers: (558,40 EURO)

- Other minor costs (including: travel expenses, 15.337,00 EURO)

- Overheads: (16736,14 EURO)

- Total costs: (78.137,36 EURO)

- Requested Contribution: (78,137.36 EURO)

- Permanent Staff costs: Professor (0.5 man-months, 7.000 EURO), 2 Senior Research Fellow (3.5 man-months, 45.000 EURO).

1.6 Tabular overview of budget and actual person-months

Tabular overview of budget and actual person-months for the whole period of project activity (years 1, 2 and 3). In the last line the figures in brackets correspond to the estimate of the total number of man-months considering the project extension from 36 to 39 months.

Contract n.	34721	Partner Person-months (PM) per Workpackage						AC - own staff			
Acronym:	TAGora										
Reporting Period:	1, 2 and 3										
		TOTALS	PHYS-	SONY-CSL	UNI	UNIK	UNI-	**AC**	UNI	UNIK	UNI-
			SAPIENZA		KO-LD		SOTON	**TOTAL**	KO-LD		SOTON
Workpackage 1	**Actual PM:**	50.4	17.5	7.4	3.5	7	15	0	0	3	8
Emergent Metadata	**Planned PM:**	46	15	9	3	3	16	0	0	3	7
Workpackage 2	**Actual PM:**	60.7	0	23.5	17.7	18	1.5	0	0.25	3.25	
Applications	**Planned PM:**	48	0	20	13	15	0	0	0.25	4	0
Workpackage 3											
Data Analysis or emer-	**Actual PM:**	100.7	42.58	24.7	18.9	8	6.5	0	2.25	0.25	2
gent system properties	**Planned PM:**	71	34	32	15	15	4	0	2.25	1	2
Workpackage 4	**Actual PM:**	124.4	72	1.65	16.7	22	12	0	0.25	0.25	2
Modeling and simulations	**Planned PM:**	124	60	3	12	8	12	0	0.25	1	1
Workpackage 5											
Dissemination and	**Actual PM:**	62.8	17	21.3	4	14.5	6	0	0.25	0.25	
exploitation	**Planned PM:**	42	15	16	3	4	4	0	0.25	1	0
Workpackage 6	**Actual PM:**	16	16	0	0	0	0	0	0	0	
Management	**Planned PM:**	18	18	0	0	0	0	0	0	0	0
	Actual Tot.	415	165.1	78.6	60.8	69.5	41	22	3	7	12
Total Project	**Planned Tot.**	349	142	80	46	45	36	23	3	10	10
		(378)	(153.8)	(86.7)	(49.8)	(48.8)	(39)				

1.7 Cost budget follow-up table

THE "COST BUDGET FOLLOW-UP TABLE" WILL BE INSERTED BY PHYS-SAPIENZA IN THE FINAL VERSION OF THE FINAL MANAGEMENT REPORT.

1.8 Major deviations from cost and person-month budget

1.8.1 Major deviations from Cost Budget

PHYS-SAPIENZA:

SONY-CSL: In the third year, a lower number of employees at Sony CSL Paris for a period of about a year resulted in a higher per person monthly indirect fee (actual costs). In addition, we recruited two additional junior staff members in order to perform specific development work and statistical analyses. They represent an extra budget but no extra EU contribution as we agree to maintain our overall EU contribution constant.

UNI KO-LD:

UNIK:

UNI-SOTON:

1.8.2 Major deviations from Person-Month Budget

PHYS-SAPIENZA: More person months than originally planned were spent due to project extension as well to a stronger effort devoted to the realization of the feedback between theoretical approaches and modelling to control in applications.

SONY-CSL: none

UNI KO-LD: More person months than originally planned were spent due to the employment of less senior personnel. Because of a lack of experience the more junior personnel needed more hours to finish the tasks that more senior personnel would have finished quicker. This is especially true for undergraduate students who are very cheap, but also take much longer. The additional person month were evenly spend on the work packages. The bigger difference for WP4 is due to the fact that most work in WP4 was done at the end of the project.

UNIK: More person months than originally planned were spent due to project extension as well as the employment of less senior personnel. The people hired very less senior than expected. Because of a lack of experience the more junior personnel needed more hours to finish the tasks that more senior personnel would have finished quicker. This is especially true for undergraduate students who are very cheap, but also take much longer.

UNI-SOTON: The planned person-month for the last year of the project was 10 months. However, we have managed to pull in additional resources at no extra cost to the project, to help with service deployment and LSS implementation. These resources proved extremely valuable and were vital to the success of LSS, as an application as well as a dissemination activity. The additional person-month were free and hence it did effect our overall project budget.

Chapter 2

Final Form C Financial statement per activity for the whole duration of the project, to be completed by each contractor

ALL PARTNERS should provide 2 original signed paper copies of the Final Form C and of the Final Summary Financial Report. Please send these documents also electronically.

Please send an electronic copy of the Final Form C + the Final Summary (before signature) to Antonella Giampaglia (Antonietta.Giampaglia@roma1.infn.it) so that she can check it and eventually suggest modifications before the official signature.

Please be sure to complete the form with:

- Contractor's stamp
- Name of the person responsible for the work, date and signature
- Name of the Financial Officer, date and signature

Chapter 3

Summary financial report

The "Final Summary Financial Report" is a summary of the individual contractor's Form C information. It includes the total (direct and indirect) costs - as claimed by each contractor and activity type - incurred during the whole duration of the project.

This summary will be inserted by PHYS-SAPIENZA in the final version of the Final Management Report.