SIXTH FRAMEWORK PROGRAMME
PRIORITY #2
INFORMATION SOCIETY TECHNOLOGIES

IST call 5
FP6-2005-IST-5

Contract for:

SPECIFIC TARGETED RESEARCH PROJECT

Annex I – "Description of Work"

Project acronym: TAGora
Project full title: Semiotic Dynamics in online social communities
Proposal/Contract no.: 34721
Related to other Contract no.:

Date of preparation of Annex I: 28/03/06
Operative commencement date of contract:01/06/2006

1. Table of contents

1. Project summary

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

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

The project aims at exploiting the unique opportunity offered by the availability of enormous amount of data. This goal will be achieved through: (a) a systematic and rigorous gathering of data that will be made publicly available to the consortium and to the scientific community; (b) designing and implementing innovative tools and procedures for data analysis and mining; (c) constructing suitable modeling schemes which will be implemented in extensive numerical simulations. We aim in this way at providing a virtuous feedback between data collection, analysis, modeling, simulations and (whenever possible) theoretical constructions, with the final goal to understand, predict and control the Semiotic Dynamics of on line social systems.

2. Project objective(s)

To successfully navigate one’s way in a sea of information, one needs a system of fix points, a coordinate system and maps. In most current day large-scale information systems (e.g. enterprise-resource planning systems, knowledge management systems) the coordinate system is given by some kind of ontology, the fix points are given by annotations that tag documents or other data and the map is produced by connecting ontology and annotations.

Unfortunately, this system becomes less and less workable as it concerns ever larger-scale information systems. On the one hand, the top down architecture of the system does not respond flexibly enough to the needs of the users that plow the sea of information. On the other, anarchical approaches mostly do not work either, because without organization of ontologies and annotations one ends up with plenty of information systems that live completely independently from each other and cannot be joint for useful exploitation.

For instance the distributed creation of taxonomies and tags and the multiplicity of conceptual schemata generate the well known problem of semantic interoperability. One solution is to standardise. The different users of a collective information system could all agree a priori to use the same taxonomies to structure their data and to use the same conceptual schemata for their data and meta-data. The tags in the owner taxonomies can then act as a shared communication protocol between peers. Unfortunately such a standardisation approach is unlikely to work for truly open-ended collective information systems in rapidly changing domains like music file sharing, picture exchange, medical imaging, scientific papers, etc.

Alternatively, it is possible that each peer has its own local taxonomy, but that these are translated into a (more) global taxonomy which is used for querying and information exchange and thus acts as an interlingua between peers. The translation to conceptual schemata of each peer could be aided by mediators [1] and achieved through automated schema matching based on finding structural similarities between schemas (see the survey in [2]). A promising recent variant of automated schema matching is based on ostensive interactions, in which agents send each other examples of the instances of schema elements so that the mapping can be made [3]. The difficulty with this approach is that a one-to-one mapping of taxonomies or conceptual schemata is not always possible. In these cases data semantics must be taken into account.

The first approach which is trying to do this is currently being explored by the Semantic Web initiative [4] and by advocates of CYC or Wordnet [5]. The data is associated with descriptions with a formal semantics, defined in terms of ontologies [6]. This approach is clearly highly valuable for closed domains, but there are known limitations when applied to open-ended information systems [7], [8], [9]. The ontologies do not capture the grounded semantics, they only constrain inference. Moreover the semantic web requires standardisation based on universal (or at least domain-wide) ontologies. But it is hard to imagine that a world-wide consensus is reachable and enforceable in every domain of human activity for which information systems are currently in use. Even in restricted domains this is hard because of an increasingly interconnected global world. Human activity and the information systems built for them are open systems. They cannot be defined once and for all but must be adaptable to new needs.

While the Semantic Web community currently focuses, eight years after the first statement of the Semantic Web vision by Tim Berners-Lee [10, 11], on theoretical issues like the definition of the adequate knowledge representation formalism, and is still looking for large-scale applications, social bookmark systems are finding increasingly large user communities on the Web in a very short time frame and a new paradigm quickly gaining impact: Folksonomies. In applications like Flickr, Connotea, Citeulike, Delicious, etc. people do not “annotate” by hyperlinks, but by terminology. The terminology used in these applications can be freely chosen, but it evolves and leads towards patterns of terminology usage in these communities and in subgroups of these communities. Hence, one observes the emergence of terminologies providing a navigation system through a large, complex and heterogeneous sea of information. This phenomenon indicates a currently ongoing grass-root creation of knowledge spaces on the Web which is closely in line with "the 2010 goals of the European Union of bringing IST applications and services to everyone, every home, every school and to all businesses" [12].

In this framework a recent point of view which is gaining increasing consensus is that of Semiotic Dynamics which studies how populations of humans or agents can establish and share semiotic systems, typically driven by their use in communication. Semiotics studies the relation between conventionalised representations (e.g. language, gestures, pictures), the conceptualisations expressed by these representations, and the real world in which the conceptualisations are grounded. Linguistics, as a subfield of semiotics, focuses more specifically on language. A language is beyond doubt the most complex form of a conventionalised representation system that is shared by large groups of human users, and its collective dimension is more important than for other forms of representations (like visuals). Recently, emphasis has been shifting from viewing language as a static body of rules that are uniformly shared by all members of a language community towards viewing language as a complex adaptive system that arises through self-organisation out of local interactions, and is constantly shaped and reshaped by language users in order to maximise communicative success and expressive power while at the same time minimising effort. This viewpoint has been adopted both by researchers empirically studying human semiotic interactions in dialog, and by researchers interested in language change and the origins of linguistic form.

The already mentioned new developments in Information Technology (such as flickr or deli.cio.us) fall precisely in this perspective and they can be seen as examples of Semiotic Dynamics at play. For instance social tagging sites, through which tens of thousands of web users share information by tagging items like pictures or websites and thus develop folksonomies, exhibit the same dynamics as observed in human languages, such as a struggle between competing tags until one dominates. As a second example there are information systems with emergent semantics, in which tags or other forms of meta-data are related to the actual data through signal or image processing algorithms. Other examples of new forms of Semiotic Dynamics occur in grounded interactions among robots whose communication systems are not pre-programmed but have to evolve in situated embodied interactions.

Semiotic Dynamics is a new field, still in its infancy, which tries to give a scientific foundation to these developments. It studies how semiotic relations can originate, spread, and evolve over time in populations. It combines recent advances in cognitive science with advances in complex systems. Cognitive science helps to provide insights into the grounding of conceptualisations through perception in the world, the highly complex cognitive processes that are implicated in the production and interpretation of representations, the processes underlying the invention and learning of concepts and their representation, and the way humans align concepts and language to co-ordinate their semiotic systems. Complex systems science is a precious aid to model and understand the collective dynamics whereby conventions can spread in a population, how conceptual and linguistic coherence may arise through self-organization or evolution, and how concept formation and expression may interact to co-ordinate semiotic systems of individuals.

The past decade has seen the first important contributions in the study of Semiotic Dynamics but most of the key questions remain unanswered. In particular the main question can be cast as follows: how do agents establish symbolic conventions between meaning and form (or sign) when there is no global control nor any prior (innate) convention, and only peer-to-peer interaction? This project moves precisely in this direction and it draws a certain number of important steps to reach a global understanding of these phenomena. The first step (WP1) aims at providing the consortium with the *raw data* needed for the subsequent work of analysis and modeling. By "raw data" we mean the emergent metadata that arise because of agent interactions in online social communities. Several online communities are readily accessible over the web: for a selected set of these systems, tools will be developed and deployed to harvest the relevant data, metadata and temporal dynamics, and to store the acquired information in a form amenable for data analysis. In order to have privileged and controllable data sources for the collaboration, we also plan to design and deploy new systems - both online systems and actual demonstrations/experiments - for the specific purpose of data collection. The next step (WP3) is that of the analysis of those data. Due to the complex nature of the systems considered, the data analysis of their emergent properties is a highly non-trivial task, which requests the contribution of specific concepts, methods and analysis tools coming from several disciplines, as different as Probability Theory, Time Series Analysis, Graph Theory, Social Network Theory, Information Theory, Clustering Analysis. The further step (WP4) is concerned with the crucial issue of modeling and simulations. A deep understanding of the observed phenomenology can only be reached by the introduction of suitable modeling, whose outcomes should be compared with the experimental findings. In this way we intend to produce a feedback cycle where the numerical simulations and the theoretical predictions can suggest new paths for the exploration of these phenomenologies and the analysis of the raw data. At the same time the new experimental findings can suggest new theoretical constructions and modeling and so on. We plan in particular a modeling activity at different scales. On the one hand it will be important to construct microscopic models of communicating agents performing language games without any central control. At a different scale we shall consider more coarse-grained probabilistic models. In both cases the dynamical rules that govern the model should be extremely simple, possibly the most basic rules one can imagine, and – most importantly - they can be readily implemented both in simulations and in actual web-based interaction systems and thus used as the foundation for new technologies. The simplicity of the modeling, provided it does not prevent the system from developing an extremely rich phenomenology, is crucial to allow for a in-depth investigation using statistical physics and complex science concepts, possibly leading to analytical approaches.

An important point to stress is that there are currently virtually no scientific publications about folksonomy-based web collaboration systems. Among the rare exceptions are [13] and [14] who provide good overviews of social bookmarking tools with special emphasis on folksonomies, and [15] who discusses strengths and limitations of folksonomies. Our project aims at significantly contributing here.

Major expected results

- Prepare a White Paper detailing the scientific challenges of the project.
- Develop tools and deploy data collection infrastructures (software, servers, network connectivity) suitable for gathering data on line social communities.
- Devise methods and algorithms for analysing raw data collected in on line social communities.
- Develop suitable modeling and theoretical constructions to understand, predict and control the Semiotic Dynamics of on line social systems.
- Develop and make publicly available innovative applications embodying novel navigation and control concepts.
- Foster the growth of new web communities revolving around the applications developed by the Consortium.
- Create the first extensive and comprehensive body of data on web-based tagging and make it available to the broader IT and scientific community.

REFERENCES

[1] Wiederhold, G. (1992) Mediators in the Architecture of Future Information Systems In: IEEE Computer, March 1992, pages 38-49. http://wwwdb. stanford.edu/pub/gio/1991/afis.ps

[2] Rahm, E., and Philip A. Bernstein (2001) A Survey of Approaches to Automatic Schema Matching VLDB Journal: Very Large Data Bases. 10: 334-350 http://citeseer.ist.psu.edu/rahm01survey.html
[3] Tzitzikas, Y. and Meghini, C. (2003) Ostensive Automatic Schema Mapping for Taxonomy-based Peer-to-Peer Systems. Proc. of CIA-2003, the Seventh International Workshop on Cooperative Information Agents - Intelligent Agents for the Internet and Web. Lecture Notes in Artificial Intelligence n. 2782, pages 78-92. August 2003 http://www.csi.forth.gr/ tzitzik/publications/Tzitzikas CIA 2003.pdf
[4] Berners-Lee, T., J. Hendler, and O. Lassila. (2001) The Semantic Web. Scientific American. May 2001.
[5] Lenat, D., George A. Miller and T. Yokoi. "CYC, WordNet and EDR - critiques and responses - discussion." In:Communications of the ACM 38 (11), November 1995, pp. 45-48. http://www.acm.org/pubs/articles/journals/cacm/1995-38-11/p24-lenat/p45-lenat.pdf
[6] Davies, J., Fensel, D., & Harmelen, F. van. (2003). Towards the semantic web: Ontology driven knowledge management. Chicester, UK: John Wiley & Sons
[7] Aberer, K.,et.al. (2003) Emergent Semantics. Principles and Issues. To appear in Proc. of the International Conference on Semantics of a Networked World. www.ipsi.fraunhofer.de/ risse/pub/P2004-01.pdf
[8] Staab, S. (2002) Emergent Semantics. IEEE Intelligent Systems. pp. 78-86. www.cwi.nl/ media/publications/nack-ieee-intsys-2002.pdf
[9] L. Steels, "The Origins of Ontologies and Communication Conventions in Multi-Agent Systems," Autonomous Agents and Multi-Agent Systems, vol. 1, no. 1, Oct. 1998, pp. 169-194. http://www3.isrl.uiuc.edu/ junwang4/langev/localcopy/pdf/steels98theOrigins.pdf
[10] Berners-Lee, T.: Realising the full potential of the Web, 1997. http://www.w3.org/1998/02/ Potential.html
[11] Berners-Lee, T.: A roadmap to the Semantic Web, 1998.http://www.w3.org/DesignIssues/Semantic.html
[12] European Commission: IST 2003-2004 Work Programme. http://www.cordis.lu/ist/workprogramme/ fp6_workprogramme.htm
[13] Hammond, T. et al.: Social bookmarking tools (I): A general review. D-Lib Magazine, 2005, 11 (4). http://www.dlib.org/dlib/april05/hammond/04hammond.html.
[14] Lund, B. et al.: Social bookmarking tools (II): A case study - Connotea. D-Lib Magazine, 2005, 11 (4).http://www.dlib.org/dlib/april05/lund/ 04lund.html
[15] Mathes, A.: Folksonomies: cooperative classification and communication through shared metadata. Graduate School of Library and Information Science, University of Illinois Urbana-Champaign, 2004 http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html

3. Participant list

Partic. Role *	Partic. no.	Participant name	Participant org. short name	Country	Date enter project**	Date exit project**
CO	1	La Sapienza University	PHYS-SAPIENZA	Italy	1	36
	2	Sony Computer Science Laboratory	SONY-CSL	France	1	36

	3	University of Koblenz-Landau	UNI KO-LD	Germany	1	36
	4	University of Kassel	UNIK	Germany	1	36
	5	University of Southampton	UNI-SOTON	UK	1	36

*CO = Coordinator
CR = Contractor
** Normally insert "month 1 (start of project)" and "month n (end of project)"
These columns are needed for possible later contract revisions caused by joining/leaving participants

4. Relevance to the objectives of the specific programme and/or thematic priority

This project aims to contribute to the economic development of the Community by advancing the state of the art in Complexity Science as relevant to IT and by exploring highly novel IT applications based on Science of Complexity. In particular, the project takes at its starting point that there has been a massive increase in the amount of autonomy and information flow and in the number of information nodes that are participating in IT processes. Moreover there is a pressing need that information systems become ever more adaptive to user needs and rapidly expanding infrastructures. Consequently, there is a much higher interdependence of processing nodes than in the past and various properties observed in complex systems (such as self-organised criticality) are now also observed in information systems. In particular, the focus of this proposal bears direct relevance to fundamental and applicative problems that are currently attracting the attention of researchers from different fields, ranging from multi-agent systems to knowledge-management, from the semantic web to peer-to-peer content distribution. This is possible because of two general traits explicitly addressed and investigated by this proposal:

- **The role of metadata.** The widespread use of electronic content and the availability of high-speed networking have created the opportunity for unprecedented levels of information sharing, at the same time bringing forth the problem of managing such a vast amount of information in ways that allow its effective and efficient use. The traditional approach of top-down, centralized categorization becomes challenged both in scalability and effectiveness, and gives way to new paradigms characterized by bottom-up, distributed, intrinsically collaborative categorizations, hinged on *emergent metadata* . In these systems - for example folksonomies - *metadata* lose the simple role of passive "labels" and take on a new role as evolving entities, displaying their own (semiotic) dynamics. That is, the complex dynamics of the system arising from the interaction of its many parts leads to hierarchies and emergent dynamical features.

- **Online communities.** The widespread use of the Internet prompted the creation of new models of user interaction, mediated by the concept of a "virtual community". As the supporting IT infrastructure evolves, the metaphor of the "virtual community" becomes more and more sophisticated, and online social networks acquire many of the *complex* aspects pertaining to the structure and dynamics of actual social networks.

Today, the *interplay of Semiotic Dynamics and online communities* is creating novel collaborative scenarios (folksonomies, social tagging, semantic approaches to peer-to-peer systems,) that are truly "complex". That is, these systems exhibit emergent properties that challenge our ability to explain them in terms of the underlying behavior of their elementary components. Two decades ago, physical systems exhibiting the same kind of challenging behaviors prompted the development of a new branch of Physics known as the "Physics of Complex Systems". Since then, the science of complexity has evolved and has developed several analytical and theoretical tools to model and understand "complex" systems. Nowadays, our technological systems have acquired comparable levels of complexity and we believe that the time has come to approach them by using the same tools developed within Physics. We believe that a unique window of opportunity is opening right now to leverage synergies between IT, computer science and the science of complexity, the unifying paradigm being the idea of emergent dynamics in a system composed by many interacting "agents". This justifies the composition of our collaboration, with members having different but complementary areas of expertise. Such complementarity must not mislead - this proposal is focused on the *Semiotic Dynamics* of *emergent metadata* in online social communities. Concept and tools from various areas of expertise will be employed to study such dynamics, but the problem itself is well defined and very hot at the moment. Indeed, given the ever increasing deployment of collaborative tagging and resource sharing technologies, both on the Internet and in corporate environments, any fundamental breakthrough in understand and/or controlling the emergent properties of these complex systems will have a significant technological impact.

The research and development activities of the present program fit very naturally the two main IST priorities, as outlined in the call for FP6/STREP projects:

- The initial phase of the project will deal with collecting actual data (WP1) from existing, live systems and analyzing them with a variety of formal tools (WP3), eventually **inferring models** (WP4) that are able to capture the essential features of the emergent dynamics, and explain how they might arise from the interactions of single agents.

- The inferred models of the emergent dynamics will be subsequently used to develop simulations (WP4) that will allow the **formulation of design strategies** targeted at attaining a specific global behavior.

In other words, we plan to move from "understanding complexity" in IT systems, to "taming complexity", i.e. controlling - to some extent - the emergent behavior. While pursuing the goals of this program, we will touch most of the focus areas of the 5th FP6/STREP call, according to the following:

- *Reconstruction of system models from incomplete sets of data* - the fact that most of the systems under study are continuously evolving in time and are only partially accessible makes it impossible to gather a complete and consistent dataset. As part of WP3 and WP4, an effort will be made to use the unavoidably incomplete experimental data of WP1 to create verifiable models of folksonomies or peer-to-peer systems. On the other hand, the "clean room" experiments of WP1 are specifically designed to provide us with complete datasets, so that we can use them as "test beds" to verify our ability of reconstructing system properties and model parameters.

- *Multi-scale simulations* - all the systems under study are intrinsically multi-scale (e.g. folksonomies), and simulation involving different levels of detail will have to be developed to validate models and assess control strategies, with techniques ranging from the microscopic approach of agent-based simulation, to stochastic models of coarse-grained observables, to model-specific simulations of the emergent dynamics.

Within the specific scope of this project, linking simulations at different scales with one another will be one of the challenges to address. Indeed, simulation will be the tool enabling us to make contact between experimental data, theoretical models and control strategies.

- *Simulation in presence of uncertainty* - because of the asynchronous and distributed nature of the agents' activity (both in live systems and in simulation), the low-level behavior of the system is intrinsically "noisy", prompting for the adoption of probabilistic models. This, in turn, will force us to introduce uncertainly and stochasticity as basic ingredients of our simulation approaches.

5. Potential Impact

Strategic Impact

Brought to its full potential, we expect the impact of TAGora may be compared to what has been achieved by Google. The reason is that the problems encountered here are (i) general, (ii) urgent, and (iii) not understood well. Right now is the right time to invest into these investigations, because folksonomies quickly gain prominence and the first mover will achieve tremendous impact in the whole knowledge technologies industries. The society that first exploits folksonomies at their best, will gain considerable return on investment in their industries, hence attract new capital and better exploit the potential of knowledge lying now dormant. Hence, EU should now invest into the foundations that allow for full exploitation of folksonomies and that bring them to their ultimate potential.

The potential impact of this project can be seen at three levels: scientific, technological, social and industrial.

Scientific Impact

Research results. The project includes partners which have a proven record of excellence in developing and publishing research results in the scientific areas covered by the project. We expect that the project will contribute significantly to provide a solid and systematic scientific foundation for the issues raised in Semiotic Dynamics and the track record of scientists involved guarantees that these research results will effectively be disseminated to the research community at large. The limited number of partners and available resources mean that effort will have to be targeted to very concrete problems, but it is expected that publication of results acts as a catalyst and lays the foundation for much additional research.

Fostering interaction between Complex Systems and IT. One of the primary objectives of this project is to bring together two communities: researchers in various domains of complex systems and researchers in various areas of Information Technology (computer science, web technologies, ubiquitous computing), in particular those facing the challenge of Semiotic Dynamics in on line social communities The goal of this project is neither to solve all the problems that exist in this context nor to exhaust the possible areas of complex systems research that might be brought to bear on them; it is rather to start a process of mutual interaction between the disciplines through concrete test cases. We want to avoid at all costs that theoretical results are developed without any relevance to the concrete problems facing the design and implementation of dynamically evolving on line social communities, and at the same time that design and engineering is purely based on intuition rather than disciplined science.

The impact that we seek on the scientific community is potentially enormous because it goes beyond the specific scientific objectives and technologies focused on in this project. We want to foster a general movement towards the interrelation of complex systems and Information Technology by showing successful examples of cooperation and by posing and solving concrete non-trivial problems. It is only by seeing clear examples that we expect the scientific research community to follow suit. These clear successful research examples will be documented and presented through regular scientific communication channels so that they are accessible to the community at large.

Human Resources. Europe is traditionally very strong in complex systems research, particularly because research in physics and biology has enjoyed very strong funding for the past decades so that there is a network of laboratories, well-organised institutes, summer schools, etc. On the other hand, research in computer science is much weaker due to decade long lack of funding in many European countries, heavy educational loads for limited teaching staff, and consequently a general lack of human resources. There is no network of computer science laboratories that could be tapped into for promoting use of complex systems science into IT. The problem is even more serious because potentially good researchers are absorbed quickly by industry to solve the pressing problems of keeping the Information Technology required by society and industry operational and expanding.

The project hopes to help deal with the human resources issue by putting a large emphasis on dissemination, particularly with respect to female researchers (see section on gender issues). The results of the project will be fed directly into academic courses at partner institutions. Educational material will be developed and made freely available. Specific initiatives will be taken, most often in the context of existing conferences or summer schools so that we can reach all European researchers, including those not directly participating in European Union funded projects.

Technological Impact

The project will design and examine in great detail a number of technologies that are essential for the development of new forms of online distributed collaboration platforms and semantic-aware applications. These issues must be studied with great care before they are applied on a grand scale. Often mechanisms work in simulation but cannot be deployed in real-world applications because too many unreasonable assumptions have been made.

We expect however that this project will contribute in a very significant way to develop the technological basis for a totally new generation of on line social communities. The results that come out from these developments will be disseminated as widely as possible so that the maximum number of IT actors can benefit.

Social Impact

Because of the high level of exposure of web-based activity, we plan to impact the very phenomenon we are studying by developing and deploying a new generation of enriched collaborative social platforms which could reshape the social approach to sharing online information. Moreover, through the web-portal we are going to develop in the project, we expect to establish a framework where experts, scholars and simple users can meet and share and develop new ideas.

Industrial Impact

Research on the evolution of communication and language has not yet reached a level of maturity so that there is a widespread industrial impact, although the project believes strongly that this impact will come very quickly if it can be demonstrated beyond doubt that very strong and novel results can be achieved.
To ensure that this impact is present, innovation actions have been incorporated that target the identification of application areas where the results of the project can make a significant difference. The applications and products will not be developed in this project itself, but we intend to realize a large diffusion of the results of these investigations, particularly towards SMEs who might benefit from them. IBM, for instance, announces experiments with folksonomies in their intranet, because the currently used taxonomy is too expensive to be maintained [Bu05a]. Microsoft also intends to invest in this area [Bu05b].
[Bu05a] Bud: IBM's Intranet and Folksonomy. In: The Community Engine. March 2005, 6. http://thecommunityengine.com/home/archives/2005/03/ibms_intranet_a.html.
[Bu05b] Bud: xFolk: An xhtml microformat for folksonomy. In: The Community Engine. March 2005, 22. http://thecommunityengine.com/home/archives/2005/03/xfolk_an_xhtml.html

Innovation, exploitation, and dissemination

The project is highly innovative because it uses very advanced technologies in ubiquitous computing to develop and demonstrate the fundamental research emanating from this project. Exploitation is guaranteed because of the presence of educators that have a record of excellence in disseminating scientific and engineering knowledge and because of industrial partners in the project that have a record of excellence in innovation and/or simulation of spin-off companies based on technology transfer. Also awareness is guaranteed because of the educational activities at partner institutions as well as public demonstrations that will be given as part of this project in scientific exhibitions or activities oriented to the public at large such as science museums and television programs.

Added value and international cooperation

The project is embedded through its partners in various ongoing initiatives at the national and international levels. Some of its members participate in the Exystence Network of Excellence initiative which is providing a much broader umbrella for integrating the complex systems community with IT research. Also, in particular EU member countries, initiatives are being taken to stimulate the introduction of complex systems research topics into both basic research and IT applications, and some of the partners of the project are involved in them. One example is France, where CNRS has just launched a major initiative for connecting complex systems research with research in the human and social sciences. Another example is

Italy where the new-born CNR- Institute of Complex Systems (CNR-ISC), directed by Prof. Luciano Pietronero (belonging to the coordinator' team), has gathered fresh complex system resources devoted to the study of highly interdisciplinary issues.

Finally, some of the partners in the project have strong connections with research groups in Japan and in the United States so that the project can be guaranteed strong connections also with other groups working on complexity.

5.1 Contributions to standards

The interoperability between networked emergent semantics applications requires the definition of a well-defined protocol and exchange format. This format can be built on top of or as extensions to existing Web standards such as RSS, Atom, and RDF. The development and experimentation with emergent semantics software will be instrumental to test the usability or to propose amendments to the existing standards. This aspect is especially taken into account by the introduction of simple microscopic models which, once checked their effectiveness in reproducing features of real systems, could be readily implemented in actual web-based interaction systems and thus used as the foundation for new technologies.

6. Project management and exploitation/dissemination plans

6.1 Project management

The management structure of the consortium includes the Governing Committee, the Project Coordinator and Contractors.

The **Project Coordinator** will be responsible for the day-to-day co-ordination of the project, and will be the main interface between the project and the European Commission (with the help of the Governing Committee, see below). It will communicate all information in connection with the Project to the Commission, allocate the financial contribution received from the Commission to the Contractors according to the decisions taken by the Consortium, prepare annual accounts such that it is possible, at any time, if requested by the Commission or by the Contractors, to inform them of the distribution of funds among the Contractors, specifically the amounts allocated and the dates of payment to each Contractor.

In addition he/she will make sure that the deadline, structure, and content of the deliverables prepared by the contractors are in line with what indicated in the contract, address the Project Deliverables to the Commission, after prior validation by the Governing Committee.

The **Governing Committee** is the Consortium's decision-making and arbitration body. The Governing Committee will consist of one researcher from each team appointed by the lead contractor of the corresponding participant. Any expert or qualified person may be invited to attend meetings of the Governing Board in an advisory capacity. The chairperson of the Committee, who is elected by the members from among the members themselves of the

Governing Board, sets the agenda for meetings of the Governing Board and monitors the implementation of the decisions taken by the Governing Board. The Chairperson shall also convene meetings at any time: upon written request of any Contractor in case of an emergency situation, upon request of 1/3 of the Contractors. The secretariat of the Governing Board shall be ensured by the Coordinator. It is anticipated that the Governing Committee will meet during the course of the project at approximately equal intervals, but the frequency may change according to circumstances. The Governing Committee will decide, inter alias, on the following matters:

- political and strategic orientation of the Project;
- the Consortium's budget and the financial allocation of the EU's contribution between the various activities on the one hand, and between the various Contractors on the other;
- modifications to the "Programme of Activities", including, but not limited to, decisions to abandon a research programme or to reduce the budget allocated to it;
- the Governing Committee is the decision-making body for any issue concerning the proper operation of the Consortium;
- monitor the preparation of the Project Deliverables for the Commission;
- decide on Intellectual Property Rights;
- in case of default by a Contractor, review and prepare proposals for measures to be taken against the Defaulting Contractor, including a request to the Commission for an audit, possible re-assignment of the Defaulting Contractor's tasks, staggered payment of the financial part of this Contractor's contribution, and suggestions as to any new entity to replace the Defaulting Contractor;
- in case of default by the Coordinator in the performance of his/her tasks, prepare proposals for measures to be taken and the possible appointment of a new Coordinator;
- preparation, update, and monitoring the actions that the consortium will entail for training, exploiting and disseminating the knowledge generated in the project;
- preparation, update, and monitoring of a training plan describing the actions that the consortium will realize for training the researchers hired by the consortium;
- establish and maintain the project web site;
- maximize the availability of tools and experimental platforms developed by the partners during the project;
- coordinate mobility and staff exchanges;
- coordinate synergies and cooperation with the other projects of the Complex Systems initiative;
- coordinate the organization of conferences and Workshops;

Contractors shall provide the chairperson of the Governing Committee and the Coordinator with the deliverables, information, and reports as they require in order to perform their duties under the Contract or as the Commission may request. Deliverables, information and reports shall include, inter alias, the supporting documents evidencing expenditures incurred by the Contractors for the purposes of the Project. In particular Contractors are responsible for the preparation of the deliverables indicated in the deliverables list reported in section B6 and in the updated deliverables lists that will be prepared during the course of the project in which they act as prime responsible.

Each Contractor shall address to the Coordinator an audit certificate in accordance with the relevant article of the EC Contract no later than thirty (30) days after the expiry of each certification period. For the proper management of the Project and of the Community contribution, the Contractors agree to submit to the Coordinator every twelve (12) months a cost statement of expenses incurred by each of them together with the supporting documents. Contractors shall promptly notify any delay in performance or any event that may impact the Project to the appropriate body.

Contractors acting as coordinators of the research and demonstration activities are also responsible for monitoring the corresponding activities and for informing the Coordinator of any difficulty arising in connection with the conduction of the activities.

Workgroups will be formed to co-ordinate activity within the workpackage. It is expected that most workgroup communication will be carried out electronically, although informal workgroup meetings will take place when the members are attending Management or Scientific meetings.

The day-to-day communication will be facilitated by the project information infrastructure that will be maintained by the coordination node. This communication infrastructure will include:

1. A set of WWW pages, organized as follows. There will be a home page for the whole project with links to pages describing: (a) the research themes of the project; (b) dissemination activities; (c) related work; and so on. The group leading the research efforts on a particular aspect of the project (often identified with leadership in a workpackage) will be responsible for the maintenance of the respective WWW pages.
2. A mailing list of the project partners will be used to facilitate the day-to-day communication on questions of common interest.
3. A centralised ftp area will be set up to allow easy download and upload of files by the project members.
4. A flexible, powerful and simple web-based collaboration platform (wiki-like) acting as glue for the consortium will be made available on the private section of project web site.
5. The public section of the project web site will also constitute an important dissemination tool, as described in WP5.

We will prepare a **Yearly Management Report** that will detail all management activities (not directly content-related) and cover all deliverables of management type specific to each project (e.g. setting-up websites). The management report will give a self-evaluation of the project according to the following criteria:

1. The extent to which tangible results were achieved in accordance with the overall project goal (a summary of highlights of the pas year).
2. The extent to which the project contributed to the Complex System initiative as a whole, that is contributed to developing common concepts, contributed to the complex system research working paper, help in organising common events;
3. The extent to which the project made results (data, software) available to all other projects, etc.
4. The extent to which the project contributed to making research results and research goals accessible to a wider audience including the public, industry and policy makers; the extent to which the project produces articles and presentations or other items of use in disseminating results and general goals of the complex system initiative.
5. Quantitative assessment: How many articles were published? How many in the specialised press, how many in the non-specialised press? What strategy and what actions were taken to attract the press?

5. What actions were taken to present research results to industry?
6. Content and maintenance (regular updating!) of web-site: how does it fit in the overall dissemination strategy of the initiative? Is it always up-to-date?

6.2 Plan for using and disseminating knowledge

To assure that the results of the project will be disseminated to a wider scientific and non-scientific audience, the following actions will be taken:

- 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;
- 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.
- disseminate results to the press at large on order to diffuse them as widely as possible;
- identify all the possible actions that can be taken, beside those already described in this document, to disseminate the results to the other projects of the Complex System Cluster and to favor synergies and possible joint activities;
- establishing and maintaining a project web site (see WP5 and D5.2);
- maximize the availability of tools and experimental platforms developed by the partners during the project to the other projects of the Complex System Initiative and to the scientific community as large;
- maximize the dissemination in the research community as large of the White Paper (D5.3) that will be produced during the first part of the project;
- exploit the fact that many partners are involved in dissemination and educational activities to train students (including Ph.D. students) and to develop course material that can be used to spread further the results of this project;
- 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 which 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. A special effort will be made to reach female researchers (see section on gender issues);
- all members of the project will be encouraged to publish their results, and early-stage researchers in particular will be encouraged to present their work at international conferences.
- we will acknowledge funding via the FP6-IST Complex Systems initiative in all publications and presentations.

6.3 Raising public participation and awareness

Some of the project partners have already wide experience with showing demonstrators to the public at large, particularly in the context of science museums or art/science events. For example the Talking Heads experiment developed by one of the partners was shown in the Palais de la Decouverte (Paris), the Whipple museum (Cambridge), the Wellcome Gallery (London), the Spoleto Science Festival, etc. To raise larger public awareness into the capabilities of present-day technologies and raise common interest, particularly by young people, the project will pursue these kinds of activities with the demonstrators that are developed.

Similarly the project partners will continue to disseminate their activities through the international press, as a way to reach a larger public but also as a way to inform researchers in other fields of study and technology development which would otherwise not become knowledgeable about this research.

More specifically, several applications that will be developed in the framework of this project (see WP2) are precisely hinged on creating new social on-line communities and exposing the members of these communities to innovations developed by the consortium. This kind of exposure is very broad in nature and involves the Internet at large.

7. Workplan– for whole duration of the project

7.1 Introduction - general description and milestones

The project aims at exploiting the unique opportunity offered by the availability of enormous amount of data. This goal will be achieved through: (a) a systematic and rigorous gathering of data (WP1) that will be made publicly available to the consortium and to the scientific community; (b) devise and implement specific brand-new applications (WP2); (c) designing and implementing innovative tools and procedures for data analysis and mining (WP3); (d) constructing suitable modeling schemes and theoretical constructions (WP4). We aim in this way at providing a virtuous feedback between data collection, analysis, modeling, simulations and (whenever possible) theoretical constructions, with the final goal to understand, predict and control the Semiotic Dynamics of on line social systems. The outcomes of WP2 are the central part of the project (see sect. 7.3) since the applications developed there:

(i) act as source of data (to feed WP3 and WP4);

(ii) represent the ideal platforms to check the theoretical understanding, implementing control strategies and experiment innovations;

(iii) disseminate the innovations developed by the Consortium to a broad audience raising the public awareness and creating new social on-line communities (WP5).

The main milestones of the project are set to:

(i) Month 5: Data collection can start after suitable hardware and software tools have been deployed.

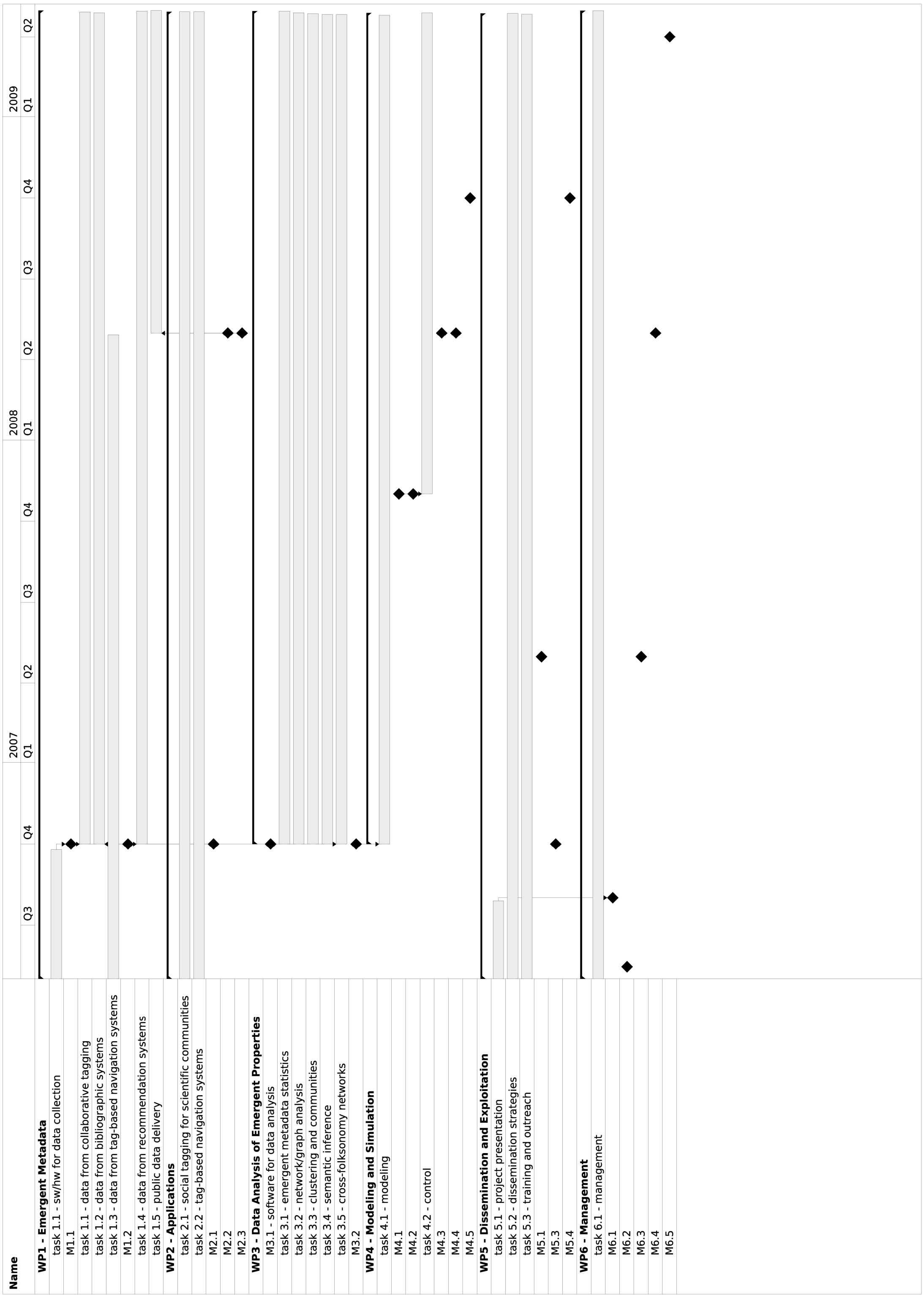
(ii) Month 5: From this point on, and in parallel with the data acquisition, data analysis and modelling activities will start.

(iii) Month 17: When a solid knowledge of the modeling and theoretical schemes suitable to understand the phenomenology of online social systems has been acquired, the prediction and control phase can follow.

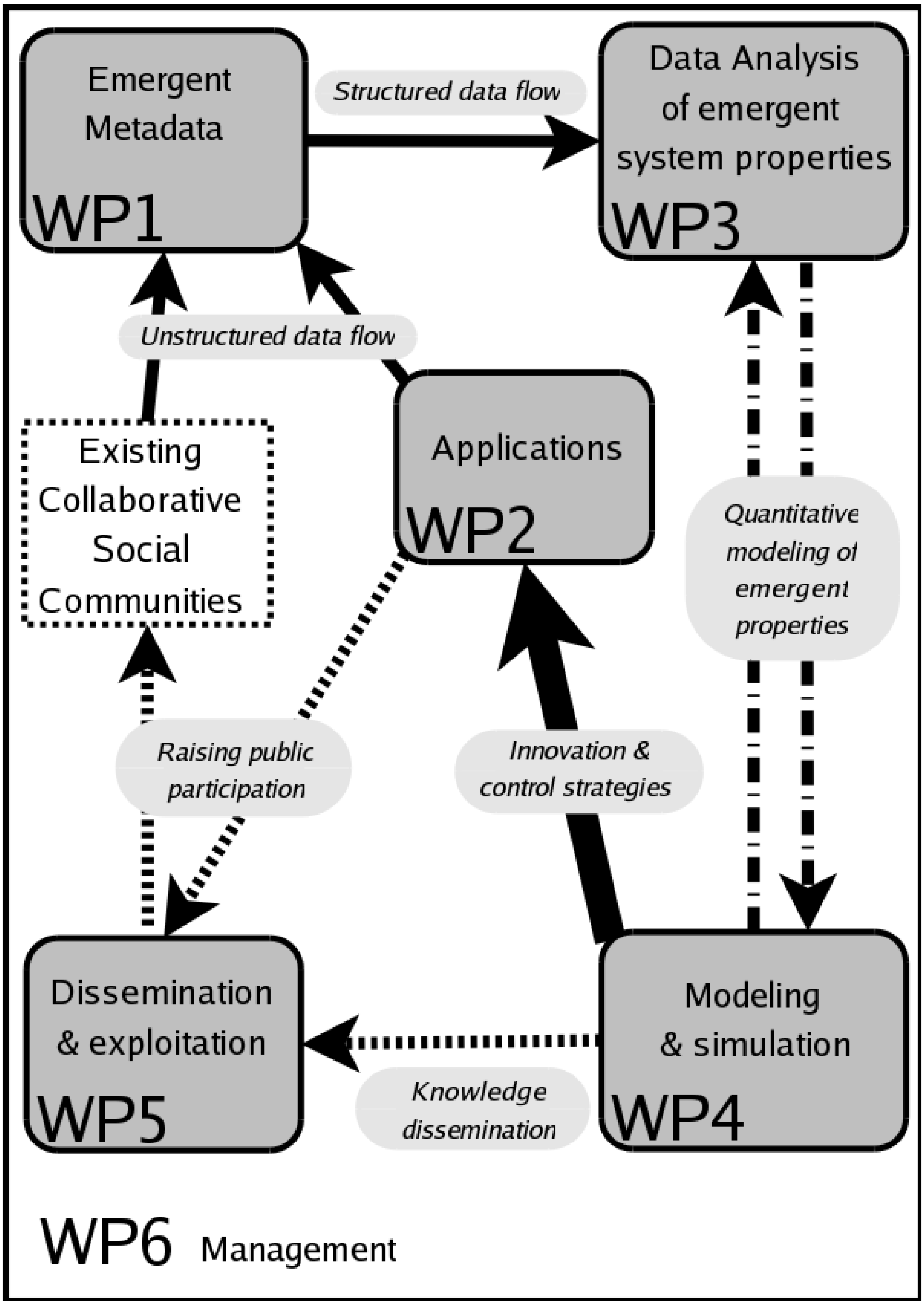
(iv) Month 23: Final versions of the applications, embodying control strategies, are delivered.

(v) Month 35: Data collected and tools developed by the Consortium will publicly released.

7.2 Workplanning and timetable



7.3 Graphical presentation of work packages


Emergent Metadata
WP1
Structured data flow
Data Analysis of emergent system properties
WP3
Unstructured data flow
Existing Collaborative Social Communities
Applications
WP2
Quantitative modeling of emergent properties
Raising public participation
Innovation & control strategies
Dissemination & exploitation
WP5
Knowledge dissemination
Modeling & simulation
WP4
WP6
Management

The above graph sketches the information flow and functional dependences among the different Tasks composing the program. The smoothed rectangles denote WP while the colored rectangles indicate the Tasks. The large white enclosing rectangle represents the relation that all Task units bear to the Management (WP6) and Dissemination (WP5) WPs.

7.4 Work package list /overview

a) Detailed implementation plan introduction
(explaining the structure of this plan and the overall methodology used to achieve the objectives)

The project is structured in such a way to have a fruitful interplay between theory, numerical simulations and experiments. Theoretical research and numerical simulations will take place in close interaction with technological investigations and there will be a very close loop between theory and experimentation, otherwise technology will be generated without proper mathematical foundation, or mathematical theories will be developed without any bearing on the practical problems involved in building embodied communicating agents.

The project is structured into six Workpackages which are all strongly interrelated and so there will be a systematic cross-investigation.

1. WorkPackage 1 - Emergent Medadata
2. WorkPackage 2 - Applications
3. WorkPackage 3 - Data analysis of emergent system properties
4. WorkPackage 4 - Modeling and Simulation
5. WorkPackage 5 - Dissemination and exploitation
6. WorkPackage 6 - Management

d) Detailed work description broken down into workpackages:

(Workpackage list, use Workpackage list form below)

Workpackage list

Work-package No[1]	Workpackage title	Lead contractor No[2]	Person-months[3]	Start month[4]	End month[5]	Deliv-erable No[6]
1	Emergent Metadata	5	46	0	35	D1.1, D1.2, D1.3, D1.4
2	Applications	4	48	0	35	D2.1, D2.2, D2.3, D2.4, D2.5
3	Data analysis of emergent system properties	3	71	5	35	D3.1, D3.2, D3.3, D3.4, D3.5
4	Modeling and simulations	1	124	5	35	D4.1, D4.2, D4.3, D4.4, D4.5, D4.6
5	Dissemination and exploitation	2	42	0	35	D5.1, D5.2, D5.3, D5.4, D5.5
6	Management	1	18	0	35	D6.1, D6.2, D6.3, D6.4
	TOTAL		349			

[1] Workpackage number: WP 1 – WP n.
[2] Number of the contractor leading the work in this workpackage.
[3] The total number of person-months allocated to each workpackage.
[4] Relative start date for the work in the specific workpackages, month 0 marking the start of the project, and all other start dates being relative to this start date.
[5] Relative end date, month 0 marking the start of the project, and all ends dates being relative to this start date.
[6] Deliverable number: Number for the deliverable(s)/result(s) mentioned in the workpackage: D1 - Dn.

7.5 Deliverables list

Deliverable No[7]	Deliverable title	Delivery date [8]	Responsible Contractors	Nature [9]	Dissemination level [10]
D1.1	(Task 1.1) – Data delivery from selected folksonomy sites (in raw and/or post-processed form)	11	1	O	CO
D1.2	(Task 1.2) - Data delivery from bibliographic reference sharing systems. (Task 1.3) - Data delivery from experimental tag-based navigation systems.	23	2, 3, 4	O	CO
D1.3	(Task 1.4) - Data delivery from selected recommender systems.	11	5	O	CO
D1.4	(Task 1.5) – Public delivery of data collected by the Consortium and related documentation.	35	5 (All)	O	PU
D2.1	(Task 2.1) - First version of social tagging system for bibliographic data. The system will allow for sharing BibTeX-based bibliographic data. It will provide easy access to the raw data for the consortium, and means to track the evolution of its content over time. (Task 2.1) - First version of folksonomy peer-to-peer system for sharing of bibliographic data. The system will support the sharing of tagged resources on top of a peer-to-peer technology. It will integrate the open source bibliography editor JabRef.	11	3, 4	D	PU
D2.2	(Task 2.2) - First version of the Tag-based navigation system for images. (Task 2.2) - First version of the Tag-based navigation system for music.	11	2	D	PU
D2.3	(Task 2.1, 2.2, 2.3) - Interim Report on tagging systems update and usage. The report will summarize the experiences gained from the running systems, and draw conclusions for the work in months 23-35.	23	4 (All)	R	PU
D2.4	(Task 2.2) - Final version of the Tag-based navigation system for images. (Task 2.2) - Final version of the Tag-based navigation system for music.	35	2	O/D	PU

[7] Deliverable numbers in order of delivery dates: D1 – Dn

[8] Month in which the deliverables will be available. Month 0 marking the start of the project, and all delivery dates being relative to this start date.

[9] Please indicate the nature of the deliverable using one of the following codes:

R = Report
P = Prototype
D = Demonstrator
O = Other

[10] Please indicate the dissemination level using one of the following codes:

PU = Public
PP = Restricted to other programme participants (including the Commission Services).
RE = Restricted to a group specified by the consortium (including the Commission Services).
CO = Confidential, only for members of the consortium (including the Commission Services).

D2.5	(Task 2.1) - Final report on tagging systems update and usage. The report describes the final version of the tools, and in particular the built-in features developed within the project.	35	4 (All)	R	PU
D3.1	Tools and report for extracting emergent metadata statistics and network metrics (Month 11). Based on data formats described in WP1, 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.	11	3 (All)	R	PU
D3.2	Methods for identifying communities. Based on results of WP1 and D3.1 these methods will describe how to identify communities in large folksonomies (cf. Task 3.3) and how to represent them in a standardized, re-usable format. In particular, these methods will include formal concept analysis tools and methods.	23	4 (All)	R	PU
D3.3	Methods for using semantic inference in data analysis. These methods will extend methods described in D3.1 and D3.2 to include semantic background knowledge. The deliverable will include an evaluation to which extent semantic inferences help to improve analysis.	23	3	R	PU
D3.4	Methods for tracking research topic emergence. These methods will report upon investigations into how tags spread between different communities and how they can be traced and predicted.	35	5	R	PU
D3.5	(Task 3.5) - Protocol for integrating cross-folksonomy networks.	23	5	O	CO
D4.1	(Task 4.1) - Review of theoretical tools for modelling and analysing Collaborative Social Tagging Systems.	11	1 (All)	R	PU
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. (Task 4.1) - Report on the roadmap leading from modeling activity to control strategies.	23	1 (All)	R	PU
D4.3	(Task 4.1) - Set of software simulators implementing the best performing modeling schemes and the ensuing control strategies.	23	1	O	PU
D4.4	(Task 4.2) - Review of existing recommendation strategies and systems.	11	5	R	PU
D4.5	(Task 4.2) - Deployment of a semantic recommender.	35	3	O	PU
D4.6	(Task 4.2) - Report describing the results of the control experiments performed.	35	1 (All)	R	PU
D5.1	(Task 5.1) - Project presentation report	4	1	R	PU
D5.2	(Task 5.2) - Project Web site for the project	4	1	O	PU
D5.3	(Task 5.3) A *White Paper* that will describe target problems and grand challenges for Semiotic Dynamics systems, clearly recognised by all and openly communicated to the scientific community.	11	1 (All)	R	PU

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.	23	1	O	PU
D5.5	(Task 5.3) Report on the impact, usability and user communities characterization of our web-based experiments and demos.	35	2	R	PU
D6.1	Provision of reports as required to the Commission.	See WP text	1	R+O (M5.1, M5.3)	PP
D6.2-D6.4	Yearly Management Report (see section 8.3)	11,23,35		R	PU

7.6 Work package descriptions

Workpackage 1 (WP1) – Emergent Metadata

Workpackage number 1	**Start date or starting event:**			**0**	
Workpackage title Emergent metadata					
Participant id	**1**	**2**	**3**	**4**	**5**
Person-months per participant:	**15**	**9**	**3**	**3**	**16**

Objectives

The objective of this work-package is to provide the Consortium with the *raw data* needed for the subsequent work of analysis and modeling. By "raw data" we mean the **emergent metadata** that arise because of agent interactions in online social communities, as described in the introduction. Several online communities are readily accessible over the web: for a selected set of these systems, tools will be developed and deployed to harvest the relevant data, metadata and temporal dynamics, and to store the acquired information in a form amenable for data analysis. The deliverables of WP1 will consist of software tools to acquire data, hardware platforms to deploy them, and collected datasets. In this logic, the milestones of WP1 will be of a purely technological nature and they will eventually benefit a broader community than the Consortium itself.

Assessment and evaluation elements

A measure of success of this workpackage will be given by the timeliness and accuracy with which the Consortium will deliver the structured data coming both from already existing on-line social systems and from the applications developed inside the Consortium (WP2). Both the timeliness and the accuracy will be crucial elements to feed the activities of data analysis (WP3) with statistically significant amount of data to discover, monitor and analyze emergent properties of online social communities.

Description of work

Since the paradigm of *emergent metadata* is applicable to a broad set of online communities, the process of data collection will take on a different form for each of the specific systems we plan to study. As a consequence, the work associated to this work-package is subdivided into tasks, each task corresponding to a different type of online social community or system, with its own specific kind of emergent metadata.

Task 1.1 - Data from collaborative tagging (folksonomies)

This task covers automated, systematic collection and storage of data from collaborative tagging systems. The output of this task consists of 1) a set of clients that are able to extract relevant data through public APIs or HTML parsing, suitably abstracting site-specific details and storing collected data in a way which is amenable to post-processing 2) a hardware infrastructure (servers/workstations/network) to deploy the data collection software.

Raw data

One of the most interesting aspect of collaborative tagging, from a researcher's point of view, is that a large amount of raw data is available for analysis, due to the fact that openness and the ability to access others' tagged resources are fundamental features of collaborative tagging itself. The information that is relevant for data analysis is system-dependent to some extent, but it broadly consists of the following categories:

- the list of tagged resources, together with some kind of unique identifier that allows to distinguish them within the system
- the set of tags associated to a given resource
- the set of users, and the social network of their relations within the system, when applicable (some systems allow "friend" or "contact" relationships among their users)
- the set of resources associated to a given tag
- the temporal series describing the past evolution of the system in terms of inserted resources, added tags, new users and so on.
- the temporal series associated with the real-time evolution of the system
- system-specific details and other relevant metadata

Access to raw data

Most of the above information is readily available when accessing the system interactively, through a browser, but it is not presented in a form which is amenable to automated processing. Systematic data collection requires the development and deployment of data collection software. To date, there is no widespread standard to expose the information contained in collaborative tagging systems. Most of such systems adopt a mix of web-based technologies in order to make their contents available to third-party applications. "Flickr" (http://flickr.com), for example, exposes all of its information through a public API based on XML-RPC and REST. Language bindings are provided for a wide variety of development environments and made available as libraries that developers can link to.
On the other hand, there are popular systems (http://del.icio.us, for example) that do not expose the full system state through a public, coherent API. In these cases, one has to resort to automated web browsing and HTML scraping/parsing, mimicking the activity of an interactive web user. Some amount of reverse-engineering is required, and care has to be taken for those cases where intense client activity can trigger denial-of-service alarms or automated throttling of the client. Finally, in some cases it might be possible to contact the site maintainers and negotiate a privileged, private access to the database, live or as a snapshot, possibly post-processed in order to protect the privacy of users.
As far as live monitoring is concerned, most folksonomy / collaborating tagging sites expose real-time evolution of the system through RSS/RDF/Atom syndication (for example recently entered resources, activity for a given tag/resource and so on). A number of client tools are readily available for automated RSS parsing, so that this is certainly a robust way to implement live monitoring of collaborative tagging, when deemed necessary.

Data collection clients and client requirements

The first deliverable of this task is a software client that is able to interface with most of the popular collaborative tagging / folksonomy sites, either by using their public APIs or through HTML parsing. Such a client will be designed in a layered fashion, so that site-specific details can be properly abstracted away and support for new sites can be easily coded in. Since the operation of the client is neither CPU-intensive nor bandwidth-intensive (due to access rate limitations imposed by most sites), performance is not an issue and the client/s can probably be coded by using a high-level scripting language like Python, Perl or Ruby. The advantage of this approach is that the low-level part of the client can leverage the large body of network libraries available for scripting languages. Python, for example, has modules that support XML-RPC and REST out of the box, allowing the developer to concentrate on the higher level aspects of client operation. Some tagging sites (http://flickr.com, for example) even provide bindings for popular scripting languages, moving the abstraction level further up. Scripting languages like Python and Perl, moreover, provide HTML parsers to be used when no comprehensive public APIs are available, and RSS/RDF/Atom parsers to access live feeds for monitoring purposes.
Client operation should support the following type of queries:

- resource-oriented queries: given a resource (or corresponding identificator), retrieve all entries corresponding to it, and the related tags and times

- tag-oriented queries: given a tag, retrieve all resources associated to that tag (and relevant metadata)
- user-oriented queries: given a user ID, retrieve all entries submitted by that user

Such types of queries should be supported for both time series acquisition and live feed monitoring. The initial set of collaborative tagging sites supported by the client should include the presently popular ones, namely http://del.icio.us, http://flickr.com, http://www.citeulike.org and the new tool deployed by the Nature Publishing Group, http://www.connotea.com/ .

In order to acquire a global snapshot of a folksonomy system, or a social network associated to it, a *crawler* should be developed that is able to explore the folksonomy in one or more of the above "directions" and integrate the resulting data to build a consistent single view of the system.

Acquired data need to be represented in a format amenable to post-processing by a variety of analysis tools. A flexible, high-level way of storing and sharing acquired data should be devised, encapsulating all the relevant data while depending as little as possible from the collection technique and the specific system under study.

Data collection hardware platform

As part of this work-package, a data collection and monitoring facility will be set up in order to deploy and run multiple instances of the data collection client, to store, post-process and re-distribute the retrieved data for further analysis. The data collection tasks are not challenging in terms of necessary resources (CPU/bandwidth/storage), so that an off-the-shelve system (for example a rack-mounted Linux server hooked up to an academic network) will cover the needs of the present work-package. Nevertheless, non-trivial post-processing tasks involving tag co-occurrence analysis, reconstruction of semantic and/or social network graphs, and possibly other advanced post-processed tasks still to be envisioned might require considerable amounts of CPU. Because of this, either a multi-processor (SMP) machine or a small cluster of servers will have to be purchased, set up and integrated into the network infrastructure.

Task 1.2 – Data collection from the bibliographic reference sharing system of the consortium

In WP2, we will set up different types of social tagging systems (cf. Task 2.1). These systems, the centralized system (cf. Task 2.1.1) and the peer-to-peer system (cf. Task 2.1.2) will allow for full access to the data. The data provided by our systems in WP2 will complement the data publicly available on the web. Their advantage is the full control we have, as well as the option to exploit them for functionality developed within WP2. For their use in the project, these data have to be transformed in the format defined in Task 1.1. As regards, Task 2.1.2 additional effort will be needed in order to provide distributed logging and collection software.

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

Sony CSL will provide data from its two test beds described in WP2, Task 2.1. Data will be made available to all partners. The data collection will follow the implementation of these two systems and thus mostly occur as a side-effect of Task 2.1. The main tasks to be performed in this WP are to package the data in a suitable form for distribution to the partners as well as to provide a technical documentation. Together, data and documentation will constitute deliverable D1.2 of the present workpackage.

Task 1.4 - Collecting data from online recommendation systems

Many e-commerce systems exist on the web that provide some sort of recommendations. One example is Last.fm (www.last.fm), which is an online folksonomy web site for listening to and managing music profiles. Users can create music profiles and tag songs, albums, and genre. Last.fm make recommendations to users on which other music they should listen to, mainly based on the profiles of those users. Last.fm monitors which music users listen to with time and uses that to create its own music charts. Official music charts, which are based on sales, can be retrieved from other sources, such as top40-charts.com. These web sites do not provide the data in a way that facilitates retrieval, and hence the data will most likely need to be harvested off the web sites. This data will be used to study how tags evolve and how these tags might influence or be influence by the general music market which can be assessed by tracking changes on official chart lists. It will also be used to investigate new recommendation techniques as described in WP4.

Firstly, work will focus on designing a data model (ontology) and a scalable repository (triple store) using the latest semantic web technology to provide the necessary flexibility and richness in representing information to facilitate applying various type analysis on the collected data. Secondly, data will be collected from these web sites and stored in a clone repository along with any temporal information that can be associated with the data.

Finally, tools will be designed to continuously and automatically update this data storage by monitoring any changes or additions to the data in the original sources.

Task 1.5 – Public delivery of documented data collected by the Consortium.

A final, important task of WP1 will consist in making the collected data available to a broader community than the Consortium itself. The relevant data will cover all the systems investigated and/or developed by the consortium, covering data from existing tagging systems, data from bibliographic sharing applications, data from the tag-based navigation systems, and data from recommendation systems. This large body of raw data will provide insight into the statistical and dynamical properties of tagging, at all the different scales involved (from the behavior of single users to emergent properties), and because of this it will impact future research and application development. In order to maximize the impact potential of the Consortium data, we foresee an effort in documenting extensively both the data themselves and the systems they come from. Raw data for each system will be made available in electronic form (as far as we are allowed to), using standard and open formats, and an accompanying document will be provided. Such a document will contain all the crucial metadata allowing third parties to understand and use the raw data. The above materials will be made available on the project's web site and/or as accompanying materials in scientific publications. As part of this Task an open content licence for released data will be selected in order to maximize the outreach while acknowledging the Complex Systems cluster effort.

References
[1] P. Haase, J. Broekstra, M. Ehrig, M. Menken, P. Mika, M. Plechawski, P. Pyszlak, B. Schnizler, R. Siebes, S. Staab, C. Tempich. Bibster – A Semantics-Based Bibliographic Peer-to-Peer System. In: Proceedings of the 3rd International Semantic Web Conference (ISWC2004), November, 2004, Hiroshima, Japan. LNCS, Springer, 2004.
[2] Bibster won the system awards of the Cooperative Information Agents workshop series in September 2005.
[3] C. Tempich, M. Ehrig, C. Fluit, P. Haase, E. L. Marti, M. Plechawski, S. Staab. XAROP - A Midterm Report in Introducing Decentralized Semantics-based Knowledge Application. In: Proc. Of PAKM 2000 - Fifth International Conference on Practical Aspects of Management. LNCS, Vienna, Austria, Dec 2-3, 2004.

Deliverables

D1.1 - (Task 1.1) Data delivery from selected folksonomy sites (Month 11).
D1.2 - (Task 1.2) Data delivery from bibliographic reference sharing systems (Month 23).
(Task 1.3) Data delivery from experimental tag-based navigation systems (Month 23).
D1.3 - (Task 1.4) Data delivery from selected recommender systems. (Month 11).
D1.4 - (Task 1.5) Public delivery of data collected by the consortium and related documentation. (Month 35).

Milestones[11] and expected result

M1.1 - (Task 1.1) Implementation of software clients and hardware infrastructure to perform data collection from folksonomy sites (Month 5).
M1.2 - (Task 1.4) Design and deploy a centralised system for storing the selected online resources (Month 5).

Workpackage 2 (WP2) – Applications

Workpackage number 2		**Start date or starting event:**		0	
Workpackage title: Applications					
Participant id	**1**	**2**	**3**	**4**	**5**
Person-months per participant:	**0**	**20**	**13**	**15**	**0**

[11] Milestones are control points at which decisions are needed; for example concerning which of several technologies will be adopted as the basis for the next phase of the project.

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 is twofold. We will provide experimental systems which are on the one hand intended to further improve navigation possibilities provided by tags, and on the other hand deliver data for the research work of the project. The first objective involves building systems that add value to existing tagging sites. One possibility is to enrich navigation based on tags by adding data analysis. The combination of data feature 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 is 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 will allow us to gain 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.

The objective of this work-package is 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 plan to design and deploy systems - both online systems and actual demonstrations/experiments - for the specific purpose of data collection. This will allow 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, not only as WP1, the milestones of WP2 will be of a technological nature and they will eventually benefit a broader community than the Consortium itself.

Assessment and evaluation elements

The main criterion for success regarding the applications is their acceptance in the tagging community. This can not only be measured in the number of users, but also in the feedback of users, such as comments they leave, or frequency of using mailing lists. Another important indication for the success of an application prototype is its integration into 3rd party applications.

Therefore, to measure the success of our applications, we will answer the following two questions: (1) Has a user community evolved? (2) Is there indication that the application gets integrated into other systems?

Description of work

Depending on the type of online community, social tagging systems will need different features. As a consequence, the work associated to this work-package is subdivided into tasks, each task corresponding to a different type of online social community or system, with its own specific kind of requirements.

The implemented systems will be made publicly available in form of web services. Provided that we will be able to attract a user community with our systems, these systems will allow us to collect data from day one (WP1) to be analysed in WP3. The systems developed in this workpackage will also constitute the basis for our modelling and simulation activities proposed in WP4, where we will study the influences of different strategies and parameters on our systems.

The main criterion for success regarding the applications is their acceptance in the tagging community. This can not only be measured in the number of users, but also in the feedback of users, such as comments they leave, or frequency of using mailing lists. Another important indication for the success of an application prototype is its integration into 3rd party applications.

Therefore, to measure the success of our applications, we will answer the following two questions: (1) Has a user community evolved? (2) Is there indication that the application gets integrated into other systems?

Impact of modeling/control (WP4, Task 4.2) on applications (WP2)

An important aspect of our project is a concrete interaction between the workpackage dealing with modeling/control activities (WP4) and the present workpackage, dealing with applications. We foresee a direct feedback of Task 4.2 (control) on the technologies developed and deployed in WP2. The modeling activities of Task 4.1 will provide us with a deeper understanding of the interplay between the low-level dynamics of the systems under study and their emergent behavior, in terms of information structuring and architecture. We expect that the models developed in Task 4.1 will have some definite predictive power. Because of this, we do expect to gain control on the emergent dynamics, attained by suitably engineering the interaction that agents have with the system and with one another. This research activity, which is the focus of Task 4.2, will feed back into WP2, in several different ways:

- The models developed in Task 4.1 will link aspects of information presentation to average user behavior and statistical features of the emergent folksonomy. This causal connection will probably allow to tune the presentation part of the applications (user interface, exposure to tags, tag auto-suggestions) in order to control the structure of the growing folksonomy and its subsequent usability (for example, by affecting the "semantic breadth" of tags, avoiding noise due to mistakes and increasing the robustness of the system to perturbations).
- Modeling will also pave the way to innovative strategies for navigating the content of tagging systems, by suggesting how correlations in the data, as well as the average user behavior, can be leveraged to improve the user experience of using the system, and the ability to find "interesting" information within the system. Successful modeling of the statistical characteristic of a folksonomy will allow to link the high-frequency tags to features grounded in the data themselves (visual features, text content and so on).
- In general, understanding the dynamics of user activity and how such dynamics is causally linked to the interactions that users have with data and among themselves, will allow to enhance the "findability" of the systems we will study, develop and deploy.

Task 2.1 - Social tagging for online scientific communities

Social tagging is especially well suited to the scientific community, because this community is used to communicating in a very structured fashion. This applies both to the structure of the social network (which is known to be tightly knit and can be studied accurately by means of citation metrics) and to the shared categorizations used by the members of the community. Both features - in principle - make the folksonomy approach quite promising, and this is why we plan to focus on folksonomies for the academia. Having full control of live systems will allow us to experiment with different models of user interaction, and to verify our insights with a fast turn-over by running our systems with adapted parameter settings. The object of this task is the development of two "social experiments" with the specific purpose of fostering and tracking the growth of two collaboratives tagging systems targeted at the scientific community. The two systems will have different architecture: one will be of a centralized web site; the other will be based on a peer-to-peer network. The design and implementation of such a social tagging systems constitutes deliverable D2.1. The system will be used for providing a very complete dataset to be used for subsequent modeling and simulation in WP4.

2.1.1 - Folksonomy web site for sharing of bibliographic data

Two folksonomy web sites specializing in social tagging of academic papers do already exist, namely CiteUlike (http://www.citeulike.org/) or the recently launched Connotea (http://www.connotea.org/). However, these two providers do not provide a full access to the data. UNIK will therefore implement a "clone" of these systems. The motivation for creating a "clone" of these systems is threefold: 1) unfettered access to the complete data, 2) the possibility of tracking the full temporal evolution of the system, from the very beginning, and 3) the possibility to directly observe the influence of different parameter settings. The latter includes the option to analyse the effect of more sophisticated knowledge representations, like hierarchically structured personomies, binary relations between resources (as known from the semantic web), or statistical relationships indicating the degree of similarities of the content of different personomies. This provides the possibility to directly observe on a large natural user group the influence of knowledge representation decisions on the structure and evolution of the social network. An existing prototype offers the basic features for social tagging, but does not yet provide the functionality of the systems above, nor does it provide exclusive functionality. For attracting a larger set of users, and thus obtaining significant benchmark data, the system has to provide extended features, as well as a scalable implementation and a component for systematic data collection. Special attention will be paid to the following aspects:

- Flexibility in the way users interact with metadata. During the lifetime of the system, it is likely that this interaction model will be changed several times in order to implement new ideas and "steer" the global dynamics of the system, i.e. enforce specific features at the emergent level.
- Ease of access to raw data, in terms of user/resource/tag relations and user network structure.
- Ability of tracking the system evolution over time, at a fine-grained level.

We plan to make the system known through standard news outlets on the web, and track its evolution as it grows in terms of both user base, entered resources and tag set.

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

Bibster [1], an award-winning [2] peer-to-peer system for sharing bibliographic data developed under the coordination of the Uni Ko-Ld team, employs a generic semantic peer-to-peer platform based on Sun's JXTA. The use of a centralized ontology for annotation purposes, however, does not allow arbitrary tags as content categories. The Semantic Exchange Architecture (SEA), a new project inspired by bibster and with more flexible content tagging capabilities, overcomes those limitations and provides means for easy sharing of tagged resources on top of a peer-to-peer technology. Within a case study, a SEA-based implementation for bibliographic data management, using arbitrarily tagged citations, will be run. An integration with the open source bibliography editor JabRef, which also provides easy information categorization with self-defined descriptors, will enable an easy collection and manipulation of test data. The dataset will be structurally similar to the citeulike clone, but tagging behaviors may differ because the peer-to-peer system works completely decentralized.

Task 2.2 – Tag-based navigation systems

Sony CSL will engage in developing two different experimental tagging systems. An experimental sharing system for music data as well as an experimental sharing system for image data will be implemented. The combination of these two systems is not only interesting because it allows us to study two different modalities. Moreover, the combination is ideal since a music system naturally represents a broad folksonomy, i.e. one data item is tagged by many users, whereas an images sharing system naturally represents a narrow folksonomy, i.e one user tags one data item. This lays the foundation to investigate different strategies and parameters in both contexts (WP4).

The experimental systems are not primarily seen as commercial systems but as test beds that will help the collaboration to collect data which is directly relevant for the present project (WP1). In order to be able to attract a sufficient number of users, it is important to offer added value over existing systems such as Flickr or Last.fm. To this end, we will implement systems that offer navigation by data features in addition to already existing navigation through tags. This will yield new links between data and thus enrich the navigation possibilities.

The relation between data features and tags will be investigated. We observe that only a fraction of tags can be grounded, e.g. low-level tags such as "red", or "blackandwhite" for images, and "loud" or "fast" for music. We expect that a simple one-to-one mapping from a tag to a category that can be described in terms of data features is only possible for these low-level tags. High-level tags such as "abandoned", "decay", tags denoting locations such as "Paris", or tags denoting persons will not show the same behaviour. We will thus investigate possibilities to achieve an indirect grounding, e.g. by exploiting the co-occurrence relation between high-level and low-level tags. The findings of these experiments will then be used to implement tag proposition into our systems, in order to assist the user in the tedious process of tagging. This mechanism will allow us to study the influence of tag suggestions on convergence properties as well as on the precision of tagging.

The implemented systems will be made publicly available in form of web services. Provided that we will be able to attract a user community with our systems, these systems will allow us to collect data from day one (WP1) to be analysed in WP3. The systems developed in this workpackage will constitute the basis for our modelling and simulation activities proposed in WP4, where we will study the influences of different strategies and parameters on our systems.

The main criterion for success regarding the applications is their acceptance in the tagging community. This can not only be measured in the number of users, but also in the feedback of users, such as comments they leave, or frequency of using mailing lists. Another important indication for the success of an application prototype is its integration into 3rd party applications.

Therefore, to measure the success of our applications, we will answer the following two questions: (1) Has a user community evolved? (2) Is there indication that the application gets integrated into other systems?

Deliverables

D2.1 - (Task 2.1) - First version of social tagging system for bibliographic data (Month 11): The system will allow for sharing BibTeX-based bibliographic data. It will provide easy access to the raw data for the consortium, and means to track the evolution of its content over time.
(Task 2.1) - First version of folksonomy peer-to-peer system for sharing of bibliographic data (Month 11): The system will support the sharing of tagged resources on top of a peer-to-peer technology. It will integrate the open source bibliography editor JabRef.
D2.2 - (Task 2.2) - First version of the Tag-based navigation system for images (Month 11).
(Task 2.2) - First version of the Tag-based navigation system for music (Month 11).

D2.3 - (Task 2.1, 2.2, 2.3) - Interim Report on tagging systems update and usage (Month 23): The report will summarize the experiences gained from the running systems, and draw conclusions for the work in months 23-35.
D2.4 - (Task 2.2) – Final version of the Tag-based navigation system for images (Month 35).
(Task 2.2) – Final version of the Tag-based navigation system for music (Month 35).
D2.5 - (Task 2.1) -Final report on tagging systems update and usage (Month 35): The report describes the final version of the tools, and in particular the built-in features developed within the project.

Milestones and expected result

M2.1 - First version of social tagging system for bibliographic data (Month 5).
M2.2 - Definition of the control strategy and decision about improvements for the final version of the system for images (Month 23).
M2.3 - Definition of the control strategy and decision about improvements for the final version of the system for music (Month 23).

Workpackage 3 (WP3) – Data analysis of emergent properties

Workpackage number	**3**	**Start date or starting event:**		**5**	
Workpackage title	**Data analysis and theoretical tools**				
Participant id	**1**	**2**	**3**	**4**	**5**
Person-months per participant:	**34**	**3**	**15**	**15**	**4**

Objectives

Examining quantitative aspects of folksonomy is a highly requested area of research. The objective of the WP is 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 is to identify and quantify emergent properties of the system in study, i.e. properties that can not be simply inferred from the behavior of the single agent. Beyond to suggest the collection of new or more refined raw data, the results of the data analysis will be 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. providing 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.

Description of work

Due to the complex nature of the systems considered, the data analysis of their emergent properties is a highly non-trivial task, which requests the contribution of specific concepts, methods and analysis tools coming from several disciplines, as different as Probability Theory, Time Series Analysis, Graph Theory, Social Network Analysis, Information Theory, Clustering Analysis. Each contribution will address an aspect of systems.

Task 3.1 - Emergent metadata statistics

The first data analysis to be performed will concern with the quantitative statistical analysis of the emergent metadata, as for instance the set of user defined tags in folksonomy.

This will surely include the frequency distribution of the metadata (tags/keywords distributions). It is easy to foresee the observation of non-Gaussian statistics, i.e. the appearance of fat-tailed distribution, similar to the power law observed in other ranking problems (see [1] for instance). Due to the dynamical nature of the systems in study, the time evolution of the distribution or of its first moments have to be explicitly considered. Examining this sort of distributions could give a better indication of whether folksonomy converges on terms and fosters consensus, if and how the vocabulary of tags grows/scales as the number of users grows, and if the distribution flattens or narrows, perhaps indicating less or more agreement.

Time evolution and time auto-correlation of the frequency of the typical or of the most popular tag surely represent interesting observables. This kind of analysis, borrowing concept and methods from Probability Theory and Time Series Analysis, will benefit and orient the modeling of systems in terms of Naming Games Model [2] and Stochastic Models (see WP4).

Task 3.2 - Network/graph analysis

A much more complete analysis of the systems is based on their network representation. The raw data from WP1 can be used to define a very complex dynamically evolving network. The network is characterized by three different types of nodes [3]: **users/actors**, **resources/instances** (web addresses, pictures, scientific citations, etc...), **tags/concepts**

Correspondingly several types of links are defined:

- **user vs. resource**: Each user links to a series of resources, so defining a *user resource list* (her/his bookmark list, her/his portfolio of images, her/his bibliography...);
- **user vs. tag**: Each user links to a list of tags, so defining a *user tag list;*
- **resource vs. tag**: Each resource in the user resource list is linked to one or more tags;

This complex (*tripartite*) network can be projected/reduced in several different ways to evidence particular features. Particular attention will be devoted to the **semantic networks of tags**.

Semantic networks: The nodes of these networks are the whole set of tags, as defined by the social community of users, and dynamically evolving in time. Links can be assigned between tags according to different rules, possibly assigning a weight to each link. Such links represent semantic associations of the "emerging language". This is an **emergent network** with a non-trivial unknown structure and dynamics. For instance one can link two tags if they are associated to the same resource. Alternatively to this **content based** semantic network, one can define a **social based** semantic network, where links are assigned if tags are used by the same user. More generally, links can be weighted according to the number of common tagged instances and/or the number of users using both tags. Beyond, and related to, the **semantic network of tags**, several other emergent networks can be defined: **social networks**: two users can be linked if the same resource or a number of resources appear in their respective resource lists, forming an emerging *content based social network*. Alternatively, two users can be linked if the same tag or a number of tags appear in their respective tag lists, forming an emerging **semantic based social network**. In some systems (i.e. Filckr) users are linked directly through a real social network (not an emerging one), i.e. a fourth type of link (i.e. **user vs. user**) is present. Such a network should be analyzed and compared with the emergent social networks. **Content networks**: two resources can be linked if they share a tag or a number of tags, forming an emerging *semantic based content network*. Alternatively, two resources can be linked if they belong to the resource lists of two or more users, forming an emergent **community based content network**.

The study the tripartite network, and the corresponding emerging networks, can benefit of several protocol of data analysis, focusing in different emergent properties: **Topological properties, Dynamical properties**, **Clustering/social properties**. Methods, concept and tools for this analysis will be borrowed from Graph Theory, Social Network Theory, Clustering Analysis, Formal Concept Analysis, Semantic Inference Theory. The results will guide and pose the base for the definition of Multi Agent Models of the systems (see WP4).

3.2.1 - Topological properties

All the networks in study emerge from a social interaction of users. Consequently they are very different from simple ordered (Euclidean) lattices. While it is clear that such networks cannot even approximately be described in terms of ordered lattice, it has recently been recognized that their topological features are very different from random graphs (see [4] for a review on the subject). The properties of these complex networks are currently subject to an intense activity of research. The emerging picture is that of an intermediate kind of graphs, between the random graphs and ordered lattices. In particular, they typically exhibit a small average distance between vertices (small-world effect), typical of random graphs, together with local clustering properties that are typical of ordered lattices. Moreover, many social networks are scale-free, in the sense that they exhibit power-law distribution of the degree (number of connection of each vertex). Gathering methods and tools from recent studies, we identify as quantities of interest, the simple and conditional degree distribution, betweenness, clustering coefficients, etc. The data analysis protocols concerning the topological properties of the networks should also define and provide specific tools and methods for graph visualization.

3.2.2 - Dynamical properties

All the networks in study are dynamically evolving in time. Analyzing the time evolution of the network could be important to understand how specific topology features can arise spontaneously. Several models have been proposed in the literature of statistical physics, proposing simple mechanisms that could explain the emergence of scale free networks. The data analysis of the dynamics of the different network could be aimed to the identification of which simple model could be successful in the description of folksonomy networks and possibly to give an estimation of model parameters.
Of particular interest could be the data coming from the new scientific collaborative systems implemented in WP1 (see task 1.2, M1.2). Such systems should provide the whole dynamical evolution of a particular folksonomy system, starting from its very beginning.

Task 3.3 - Cluster/community identification

In many cases it is found that social networks are inhomogeneous, consisting not of an undifferentiated mass of nodes, but of distinct groups or communities. Within these groups there are many links between nodes, but between groups there are fewer links, producing a structure of sparsely linked communities. Quantifying such an observation and finding communities within large networks in some automated fashion could be of considerable use. In the case of folksonomy it could be interesting to study communities of users, their dynamics and their role in the formation of the emerging metadata and networks. Many methods have been proposed in the recent literature, even if no unique or generally optimal approach seems to prevail. Specific methods have to be chosen or developed for the folksonomy networks.

We will apply clustering and community identification techniques that explicitly consider the content of the resources annotated by the social community for discovering communities within the user community of the bibliographic resource sharing system. Clustering and classification of large database of contents/resources are two important approaches to organize knowledge systems. We will in particular focus on combining techniques for Formal Concept Analysis, Social Network Analysis, and Semantic Web Mining:

Formal Concept Analysis (FCA) [5,6,7] arose during the last 20 years to a powerful collection of techniques for deriving conceptual hierarchies out of given datasets. For the task at hand, FCA has to be extended to allow for the treatment of labelled graphs.

Social network analysis [8] has a long tradition of modeling such structures as (directed or undirected) graphs and of subsequently analysing them. The research challenges are to adopt these techniques to the more complex structure of folksonomies, where the structure on the set of tags (hierarchies, groups) has to be taken into account as well as the tri-mode nature (users, tags, resources) of folksonomies. Techniques for mining the semantic web [9] can also be integrated in our scenario (see below). Recent methods based on Information Theory concepts, as entropy and complexity, could also result useful [10]. Deploying these techniques for the data analysis of the bibliographic resource sharing system will provide input to WP 3, where we will develop a theoretical model describing how the users of the system organise themselves into such communities.

Task 3.4 - Semantic Inference

When we come to data analysis about folksonomies, we may observe that many independence assumptions do not hold and that in fact data analysis could be most meaningful if done using background knowledge underlying the folksonomies. The problem is that the required background knowledge, which is mainly terminological (like "chair is a kind of furniture), does not easily fit into a statistical model. This situation is comparable to using terminological knowledge in text classification [11,12]).

In fact the statistical model may benefit from explicit representational knowledge, such as found in existing terminologies – maybe ones extracted from collaborative annotations - or using ontologies. Ontologies have gained a lot of prominence for knowledge sharing, but have so far found little entry into data analysis. Without ontologies, a lot of background knowledge cannot be accounted for in data analysis. We will include ontologies into the data analysis phase, investigating new methods that consider the granularity of targeted analysis as well as the granularity of ontology concepts (e.g. lower level concepts like `CPU' need to be treated differently from higher level concepts like `physical object'). Hence, the aim of this task is to investigate how to use background knowledge (ontologies, terminologies) in data analysis also investigating how to amalgamate background knowledge and emergence of complex features by machine learning (see [13] which describes feature emergence for the related task of ontology mapping discovery).

Task 3.5 Cross-Folksonomy Networks

Folksonomy web sites are rarely closed worlds. It is quite common for individuals to be active members of several online communities and thus one would expect certain tags to spread across such communities with time. For example one could be adding images to Flickr, bookmarking web sites with del.icio.us, creating their music preference profiles in last.fm, and tagging articles in Connotea. By continuously collecting data from such folksonomy web sites, and monitoring changes and additions, we can cross reference emerging tags between the separate communities to extend and connect their individual networks to create an Integrated Semantic Network. Tags might be the only common parameter between the various folksonomies and therefore it can be used to link the separate networks together. The integrated network can be based on various types of nodes and links as described in Task 3.2. Social Network Analysis and Graph Theory can then be applied to this overall network to, for example, recognising cross-folksonomy communities that used certain tags to describe various types of objects. Such analysis can help to identify clusters of objects that are of different kinds (e.g. an image and a document and some music in one cluster), but were given the same or related tags. Such non-homogeneous clusters are very difficult to identify when analysing folksonomy web sites in isolation due to their tendency to deal with specific type of objects. This task will therefore ensure that some of the separate folksonomy networks that will harvested within this project are connected together.

Once the networks are integrated, we will apply Network Analysis and Visualisation techniques to study how tagging evolved and spread across the various communities [14], and how the commonalities between these communities can be employed as a basis for recommendations that could go beyond what most of these folksonomies currently provide (WP4). Part of this work will focus on the design and implementation of system architecture to link some of the folksonomy networks that will be provided by the consortium (in WP1).

References
[1] Newman, M. E. J. *Power laws, Pareto distributions and Zipf's law* Contemporary Physics 46, 323-351 (2005). (http://arxiv.org/abs/cond-mat/0412004)
[2] Steels, L., Articial Life Journal 2, 319 (1995).
[3] Mika, P., *Ontologies are us: A unified model of social networks and semantics.* In Proc. Of ISWC-2005, Galway, Ireland, Springer (2005).
[4] Barabasi, A-L., *Linked: How Everything Is Connected to Everything Else and What It Means for Business, Science, and Everyday Life* Plume Books (2003).
[5] Ganter, B., Wille, R., *Formal Concept Analysis: Mathematical Foundations*, Springer, Heidelberg (1999).
[6] Wassermann, S., Faust, K., *Social Network Analysis - Methods and Applications*, New York (1994).
[7] Wille, R.: *Restructuring lattice theory: an approach based on the hierarchies of concepts*, in I. Rival (ed.) *Ordered sets* Reidel, Dordrecht-Boston, 445-470 (1982).
[8] Berendt, B., Hotho, A., Stumme, G., *Towards Semantic Web Mining*, in Horrocks, I., Hendler, J. (eds.) *The Semantic Web | ISWC 2002*, Proc. ISWC '02, LNCS, Springer, Heidelberg, 264-278 (2002).
[9] Ganter, B., Stumme, G., Wille, R., (eds.) *Formal Concept Analysis: Foundations and Applications. State of the Art*, LNAI. Springer, Heidelberg (2005).
[10] Benedetto, D., Caglioti, E., Loreto, V., *Language Trees and Zipping*, Phys. Rev. Lett. 88 048702 (2002).
[11] Hotho, A., Staab, S., Stumme, G., *Ontologies Improve Text Document Clustering*, Proceedings of the International Conference on Data Mining – ICDM-2003. IEEE Press (2003).
[12] Bloehdorn, S., Hotho, A., *Text Classification by Boosting Weak Learners based on Terms and Concepts*, ICDM 2004: 331-334 (2004).
[13] Ehrig, M., Staab, S., Sure, Y. *Bootstrapping Ontology Alignment Methods with APFEL*, in Proc. of ISWC-2005 – International Semantic Web Conference, Galway, Ireland, Springer, LNCS (2005).
[14] Harith Alani and Nicholas Gibbins and Hugh Glaser and Stephen Harris and Nigel Shadbolt. *Monitoring Research Collaborations Using Semantic Web Technologies*. Proc. 2nd European Semantic Web Conference (ESWC), Crete, Greece, pages 664-678, Springer (2005).

Deliverables

D3.1 - Tools and report for extracting emergent metadata statistics and network metrics (Month 11). Based on data formats described in WP1, 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.
D3.2 - Methods for identifying communities (Month 23). Based on results of WP1 and D3.1 these methods will describe how to identify communities in large folksonomies (cf. Task 3.3) and how to represent them in a standardized, re-usable format. In particular, these methods will include formal concept analysis tools and methods.
D3.3 - Methods for using semantic inference in data analysis (Month 23). These methods will extend methods described in D3.1 and D3.2 to include semantic background knowledge. The deliverable will include an evaluation to which extent semantic inferences help to improve analysis.
D3.4 - Methods for tracking research topic emergence (Month 35). These methods will report upon investigations into how tags spread between different communities and how they can be traced and predicted.
D3.5 - (Task 3.5) Protocol for integrating cross-folksonomy networks (Month 23).

Milestones and expected result
M3.1 - Acquisition of software tools for data analysis (month 5).
M3.2 - Identification of the key emergent features and global observables relevant for modeling (month 5).

Workpackage 4 (WP4) – Modeling and simulations

Workpackage number	**4**	**Start date or starting event:**			**5**
Workpackage title	**Modeling and simulations**				
Participant id	**1**	**2**	**3**	**4**	**5**
Person-months per participant:	**60**	**32**	**12**	**8**	**12**

Objectives

The objective of WP4 is twofold:

- *understanding complexity:* develop models that captures the essence of the emergent dynamics and explain how it might arise from the interactions of single agents;
- *taming complexity:* formulate design strategies that allow controlling the behavior of the system at the emergent level by suitably choosing the microscopic dynamics of the interacting agents.

Computer based simulations naturally comes into play as the tool of election to make contact between actual data collected in WP1 and WP3, models developed in the present WP and potential control strategies.

The objectives of WP4 are related to developing realistic models for the systems studied in WP1 and WP3 and gaining as deep an understanding as possible, by using a variety of modeling strategies involving multi-scale agent-based simulations.

The goal of this workpackage is to construct, implement and study specific modeling schemes aiming at reproducing, predict and control the emergent properties seen in the semiotic dynamics orchestrated in on-line communities. We plan in particular a modeling activity at different scales. On the one hand it will be important to construct microscopic models of communicating agents performing language games without any central control. At a different scale we shall consider more coarse-grained probabilistic models. Several models will be proposed to address specific aspects/scales of folksonomy. The models will allow computer simulation aimed at measuring emergent features to be compared with the results of WP3. The simulations should give an insight in how users select tags, what kind of categories and category structures underlying the evolving system of tags, how categories and tags are related to the objects being tagged, etc. It will also give us information on what kind of more global structures (such as the most frequent tags) can be provided to users to optimize their on-line community infrastructure The models will require components for assigning or adopting tags, categorizing data, and collective dynamics. However the approach will be to keep the models as simple as possible, identifying the minimal ingredients responsible for the emergent properties. The minimal character of the models should make a more analytical mathematical study feasible.

A possible way to tackle the complexity of the systems is to individuate different time scales, which can be separated. For instance, we expect that the dynamics of the social network of the folksonomy could be different from the time scale of the dynamics of the resources and of the tags. In this case one can, as a first approximation, propose a model of tags and/or resource dynamics based on a given, slowly evolving social network topology. This kind of assumption should be tested and corroborated as much as possible with the observations coming from the real data analysis (WP3).

Finally, the output of this WP has the potential to feed back into WP2, specifically to the live social tagging systems developed as part of, in order to experimentally verify the devised control strategies and demonstrate the technological advantage achieved by the present project.

Assessment and evaluation elements

The elements of evaluation of this workpackage are easily defined being related to the ability of the proposed theoretical and numerical schemes of reproducing, at the desired level of accuracy, the experimental observations. The added value of our approach is the possibility to have a control, rigorous, analytical or numerical, of the results in terms of the parameters of the theoretical model. This can trigger a virtuous loop where refined theories or modeling schemes can provide better agreement with the experimental data and the discrepancies can be progressively cured through a constant feedback with the real experiments. Once the agreement between theory and experiments reaches a satisfactory level the theoretical description can suggest and inspire controlling strategies and so on (WP2).

Description of work

The main aspects to be investigated with models and simulations will respect the scheme proposed in WP3. In particular different emerging properties would require different modeling approach.

Task 4.1 - Modeling

This task focuses on developing mathematical models and abstractions of real systems. The basic modelling strategies are the standard ones adopted in Complex Systems Science, namely identifying the minimal required set of basic ingredients that are able to reproduce some selected emergent feature of the observed data.
Such an approach has the huge benefit of defining an unambiguous measure of success for this task: the quantitative, controllable agreement between the output of the models (produced either by simulation or by analytical approaches) and data from experimental or real systems.

To clarify our strategy, let us consider the sample case of data extracted from social tagging systems (as coming from WP1). The quality of such a system relies in the emergence of a suitable semiotic system used for the navigation of the relevant resources. In order to quantify this measure of quality, one can define several metrics: understanding the factors that affect these metrics is one of the scientific and technological issues addressed by this project. Following the approach outlined above, one would select some important statistical properties (WP3), for example the exponent of the power law distribution describing the frequency of tags associated to a given resource. The higher this exponent, the more narrowly defined the resource is within the folksonomy, so that this could be regarded as a natural quality indicator of the folksonomy. Developing a model that correctly predicts the value of this exponent provides precious insights into the dynamics of the system. Not all modelling strategies are equivalent: simple models are more valuable because they allow to pinpoint crucial ingredients of the dynamics and as a consequence they indicate routes to achieve control (task 4.2). In the specific sample case at hand, standard models for power law distributions in Complex System Science could be used as a base for developing successful models of a folksonomy.

This process is a bona fide scientific investigation process, because on the one hand it leverages existing scientific literatures and expertise, on the other hand a significant fraction of its output consists of formal descriptions, models and scientific publications. Despite its theoretical character, this task has strong ties to both data acquisition (WP1) and analysis (WP3), and directly impacts the activities of task 4.2 (control).
Task 4.2 can indeed be regarding as the engineering counterpart of task 4.1, and these tasks are expected to feedback on to each other. Because of its central role in the project work plan, we expect this task to absorb a large amount of man-months and computing resources.

In the following we describe in greater details the theoretical tools that will be deployed for the actualing modelling work.

4.1.1 - Stochastic models

The systems that we plan to study and model exhibit emergent behaviors that arise from the interaction of a large number of elementary entities. In a folksonomy, for example, the shared categorization emerges spontaneously from the distributed, asynchronous - and in principle uncoordinated - tagging activity of web users. That is, there exist at least two scales: a high-level scale where "order" and the emergent dynamics are visible, and a low-level ("microscopic") scale dominated by the individual, uncoordinated activity of agents. At this scale, the activity of agents is seemingly random, and lends itself to be modelled by using probabilistic models. This is similar to what happens in statistical physics, where the appearance of macroscopic order and regularities is explained in terms of the noisy atomic behavior. Because of this, it is natural to use modeling approaches from statistical physics, based on probabilistic and stochastic models. In other areas involving complexity in IT systems, like the study of emergent properties in large-scale technological networks, the use of stochastic models has been greatly successful and we expect them to be similarly useful in the specific context. of this investigation.
Among the stochastic models that will be relevant to this WP, there are two classes of models that might turn especially successful:

Multiplicative stochastic processes, and specifically Polya's urn models and their variants [1]. These models

can be probably used to describe the collective probabilistic reinforcing of agent' choices, namely the fact that the process of tagging is somehow biased by the relative proportions of tags already existing in the system. There is preliminary evidence that this might be linked to the observed robustness of folksonomies [2], and we plan to characterize in detail the role of multiplicative stochastic dynamics in systems of social tagging.
Stochastic models for word frequencies in text, like Simon's model and its variants [3]. These models were originally devised to explain the experimental observation that the frequency of words in natural language exhibits surprising regularities, manifested in the so-called Zipf's law [4,5]. We have reasons to believe that such statistical regularities might be also present in the set of tags that users associated to a given resource. We plan to adapt this model to folksonomies and compare them with the probability distributions extracted from real systems.

Probabilistic mechanisms to explain power-law distributions. An extremely broad range of human activities exhibits characteristic statistical features consisting in power-law probability distributions ("fat tails") [4,5]. These distributions appear frequently in complex systems and are associated to the intrinsic multi-scale nature of their dynamics. Because of their ubiquity, we expect the analysis of WP3 to observe them in the data collected by WP1. Attempting to use such models to describe and explain the observed distributions will certainly provide insight into the fundamental mechanisms leading to self-organization and emergence in online social communities.

Stochastic models of network growth and evolution have been widely studied in the upsurge of interest that has been devoted to *complex networks* [6]. These models are likely to be very useful in describing the network structures than we will extract from the raw data, namely social networks of users, "semantic networks" of tags, networks of similarly tagged resources and so on. We expect that such networks will display the standard marks of complexity and emergence, so that stochastic growth models like the well-known *preferential attachment* scheme might prove relevant for the case at hand.

4.1.2 - Language Games

A natural modeling strategy useful to describe the semiotic dynamics emerging from folksonomy is represented by a newly introduced class of models, known as Naming Games [7,8,9]. In their original formulation [10], a multitude of agents interacts trying to bootstrap a common vocabulary for a certain number of objects present in their environment. Each agent manages personal lists of words associated to each object. Throughout pairwise peer-to-peer negotiations, agents try to find an agreement on the name of single objects. Despite the simplicity of the model, the system undergoes spontaneously a disorder/order transition without invoking neither a central control nor a fitness-based selection. We plan to propose similar models to describe the tagging activity of users in folksonomy. In this case, care should be devoted to include in the model definition specific features of the folksonomic system, as the topology of the social network underlying the agents interaction, or the dynamics and distribution of shared resources. The model should reproduce the statistical and dynamical features of the emerging metadata system observed in WP3.

Model simulations should give an insight in how users select tags, what kind of categories and category structures underlying the evolving system of tags, how categories and tags are related to the objects being tagged, etc. It should also give information on what kind of more global structures (such as the most frequent tags) can be provided to users to optimize their on-line community infrastructure. Many theoretical challenges are issued by the model. First of all it is important to determine which classes of interaction rules are able to make the ordering transition happen. Among these, then, it would be crucial selecting the simplest ones. It would then be necessary to understand how and when the transition takes place and which are the time scales involved with the process. This could lead to take a step backward to the interaction rules definition, and see if some possibly slightly more complicated protocols can lead, for instance, to a fastest convergence. Finally, analytical description of the process could be very precious. In particular it would be of the outmost importance determining whether the process always reaches convergence or there are other stable asymptotic states.

4.1.3 - Social Network Models

The topological and dynamical structure of the social network emerging from folksonomic system is also susceptible of a modelization based on simple microscopic rules. In analogy with recent studies on growing networks as diverse as hardware infrastructure or social communities [6], we will define models tailored to the intrinsically complex nature of folksonomy. The idea here is to take in account that the evolution of social contacts should be influenced by the quantity and quality of resources shared by each user, together with the "similarity" of users interests, as witnessed by their respective tag lists.

Task 4.2 - Control

The modeling activities of Task 4.1 above will provide us with a deeper understanding of the interplay between the low-level stochastic dynamics of the systems under study and the emergent dynamics they observe. Moreover, we expect that the models developed within Task 4.1, validated by the data provided by work-packages WP1 and WP3, will have some definite predictive power. Because of this, we do expect to gain some kind of control on the emergent dynamics, attained by suitably engineering the interaction that agents have with the system and with one another.
Even though the objective of control seems theoretically attainable, we plan to anchor the output of the present work-package to simulation and experimental activities on realistic and actually deployed systems, as described in the following.

4.2.1 Simulation and control on music and image sharing systems

Simulations at Sony CSL will involve two different models: (1) a model for a music sharing system, and (2) a model for an image sharing system. For both models, the simulations will be performed on two levels. First, the simulations will follow existing tagging systems such as Flickr or Last.fm. without offering additional grounding. The second level of simulations will adopt a simplified model of our experimental sharing systems described in WP2. The simplification lies in using only tags that can be grounded (e.g. "red", "blackandwhite"). This two-level approach enables to study the influence of data analysis on tagging systems. Studying music and image systems in parallel allows us to compare this influence in the different contexts of a broad and a narrow folksonomy. Our simulations will furthermore investigate the proposition of tags based on feature analysis compared to a proposition of tags simply based on the globally or locally most frequent tags. We will also study in this framework the influence of open access to all data compared to restricting the accessible data to the social network (e.g. direct contacts, as well as the contacts of the contacts). The optimal strategies and parameters found in these simulations will be adopted for our experimental file sharing systems described in WP2.

4.2.2 - Ontology learning

Ontology learning has so far ignored the role of the social dimension. Rather it was restricted to clustering of patterns in text ([13]) and to asking Google about patterns [14]. In folksonomies, however, the social groups play a prominent role when defining an ontology and correspondingly the ontology model should be extended in order to take this better into account. Once such a social embedding has been achieved the so far unidirectional model of associating groups with topics [15] can be pursued in two directions Such analysis does not only help the construction and alignments of distributed ontologies, it will also help to provide an overall map of the information landscape and layout where which kind of information will be found. To provide such a mechanism we will have to abstract knowledge from algorithms that find knowledge even when distributed among various individuals or groups cf. [15].

4.2.3 - Simulation and control on bibliographic reference sharing system

In this task, we will develop models for local user behavior, that explain the globally emerging structure of the network of users of our bibliographic reference sharing system, as discovered in WP3. We will simulate within the model the outcome of this locally organised user behavior in function of different parameters (availability/non-availability of recommender functions, group and topic detection functionality, etc.) of the central resource sharing system. Using standard measures of Social Network Analysis, we will evaluate which of the parameter settings have a positive effect on the evolvement of the social network. The most promising settings will be implemented within our bibliographic reference sharing system. Finally the prediction based on the model is checked against the real output of the system.

4.2.4 - Recommendations based on Network Analysis

Several e-Commerce systems provide some sort of recommendations to their users. These recommendations are often based on explicit and direct interactions the users make with the systems. Amazon.com for example uses a collaborative filtering technique for its recommendations, which is simply based on tracking what combinations of items users buy. In last.fm, recommendations are based on user profiles. These profiles are explicitly set by the users, or inferred through their choices of tags.

The rich and complex networks that will emerge from WP3 can be used to establish user-specific recommendations by taking into account the semantics of these networks [16]. Ego-centric analysis on these networks can be carried out at run time to identify closeness between individuals or objects [17]. Semantic-based Recommendations will be developed that will consider the type of network being analysed, the type of links incorporated in the network, the semantics of objects within the network, and any temporal information that might be associated with the data.

Semantic-based Recommendations will be tested on music and bibliography folksonomies collected in WP1. It will also be applied to the Integrated Network (WP3) to examine making recommendations across various communities (WP2). Recommendations systems sometimes are domain dependent, where there are designed to work with specific type of data. For this reason it is important to test the system with more than one type of domain.

References

[1] Feller, W. , *An Introduction to Probability Theory and its Applications, Volume 2*, Wiley, New York (1971)
[2] Golder, S. and Huberman, B. A. , *The Structure of Collaborative Tagging Systems* (http://www.hpl.hp.com/research/idl/papers/tags/)
[3] Ferrer i Cancho, R. and Servedio, V. D. P. , *Can simple models explain Zipf's law for all exponents?* , Glottometrics 11, 1 (2005)
[4] Newman, M. E. J. , *Power laws, Pareto distributions and Zipf's law*, Contemporary Physics 46, 323-351 (2005) (http://arxiv.org/abs/cond-mat/0412004)
[5] Ferrer-i-Cancho, R. and Sole, R. V. (2003) Least effort and the origins of scaling in human language. PNAS, 100:788--791.
[6] Barabasi, A-L., *Linked: How Everything Is Connected to Everything Else and What It Means for Business, Science, and Everyday Life* , Plume Books (2003)
[7] Nowak, M.A. and D.C. Krakauer (1999) The evolution of language. Proc.Natl.Acad.Scie. USA. Vol 96, pp. 8028-8033, July 1999.
[8] Steels, L. (2001) Language games for autonomous robots , IEEE Intelligent Systems 16 (5) (2001)
[9] Wagner, K., J.A. Reggia, J. Uriagereka, G.S. Wilkinson (2003) Progress in the Simulation of Emergent Communication and Language. Adaptive Behavior. Volume 11 (1): 37-69.
[10] Steels, L., Articial Life Journal 2, 319 (1995).
[11] Steels, L. and P. Hanappe (submitted) Interoperability through Emergent Semantics. A Semiotic Dynamics Approach. Submitted to Journal on Data Semantics
[12] Zils, A. and Pachet, F. Automatic Extraction of Music Descriptors from Acoustic Signals using EDS. Proceedings of the 116th AES Convention, May 2004.
[13] P. Cimiano, A. Hotho, S. Staab. Learning Concept Hierarchies from Text Corpora using Formal Concept Analysis. JAIR - Journal of AI Research. 24: 305-339, 2005.
[14] P. Cimiano, S. Staab. Learning by Googling. SIGKDD Explorations. 6(2), pp. 24-33.
[15] A. Löser, C. Tempich, B. Quilitz, W.-T. Balke, S. Staab, W. Nejdl. Searching Dynamic Communities with Personal Indices. In: Proc. of ISWC-2005 – International Semantic Web Conference, Galway, Ireland, Springer, LNCS, November 2005.
[16] Middleton, S., H. Alani, N. Shadbolt and D. D. Roure. Exploiting Synergy Between Ontologies and Recommender Systems. Semantic Web Workshop, World Wide Web Conf., (WWW'02), Hawaii, USA (2002)
[17] Alani, H., S. Dasmahapatra, K. O'Hara and N. Shadbolt. "ONTOCOPI - Using Ontology-Based Network Analysis to Identify Communities of Practice." IEEE Intelligent Systems 18(2): 18-25. (2003).

Deliverables

D4.1 - (Task 4.1) Review of theoretical tools for modelling and analysing Collaborative Social Tagging Systems (Month 11).
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).
(Task 4.1) Report on the roadmap leading from modeling activity to control strategies (Month 23).
D4.3 - (Task 4.1) Set of software simulators implementing the best performing modeling schemes and the

ensuing control strategies (Month 23).
D4.4 - (Task 4.2) Review of existing recommendation strategies and systems (Month 11).
D4.5 - (Task 4.2) Deployment of a semantic recommender (Month 35).
D4.6- (Task 4.2) Report describing the results of the control experiments performed (Month 35).

Milestones and expected result

M4.1 - (Task 4.1) Adoption of a set of models that capture the essence of the emergent behavior and describe them (qualitatively and, whenever possible, quantitatively) (Month 17).
M4.2 - (Task 4.2) Implementation of realistic simulation software aimed at the control experiments (Month 17).
M4.3 - (Task 4.1) Feedback to WP2 about the best control strategies inspired by the modeling and simulation activities (Month 23).
M4.4 - (Task 4.2) Implementation of a semantic recommender (Month 23).
M4.5 - (Task 4.2) Preliminary control experiments performed (Month 29).

Workpackage 5 (WP5) – Dissemination and exploitation

Workpackage number **5**		**Start date or starting event:**		**0**	
Workpackage title Dissemination and exploitation					
Participant id	**1**	**2**	**3**	**4**	**5**
Person-months per participant:	**15**	**16**	**3**	**4**	**4**

Objectives

The goal of this work package is to ensure that the results of this project are widely disseminated and can constitute the basis of training activities so that both the scientific and engineering community as well as the population at large may benefit from the results of the project.

Description of work

Task 5.1 Project Presentation.

This task will involve the preparation and publication of a brief project presentation in English accessible to the non-specialist, avoiding technical language, mathematical formulae, and acronyms as much as possible. It will be published via the World Wide Web and supplied in printed form to the Commission if requested.

Task 5.2 Dissemination strategies

The Governing committee will address the issue of using and disseminating knowledge through the lifetime of the project. This will include in particular monitoring the participants' actual achievements in dissemination and their plans at that time for the exploitation of their results - for the consortium as a whole, or for individual participants or groups of participants. The project will put great emphasis on publication through standard scientific and engineering communication channels. It will attempt to publish results in the best scientific journals and communicate work at top conferences. As much as possible we will use existing communication media to touch the largest possible audience.

The very nature of this project leads to an interaction with on-line social communities. This aspect will play a crucial role in disseminating the work of the consortium in a four-fold manner.

5.2.1 - Explicit dissemination activity on the web

A web site for the project (D5.1) will be set up in order to provide public access to the consortium activities to its results and its data. Moreover, since tagging is a comparatively new phenomenon on the web, an authoritative portal on the technology and its social implication is currently lacking. We will strive to create such a portal to attract the interest not only of technology experts, but also people from the social science community, the information society community, statistical physicists, and in general - at an introductory level –

the general public on the web.

At the same time, the members of the project will engage in fine-grained dissemination activities through existing IT portals, collaborative news outlets, dedicated blogs, and any forms that constitute “conversation” on the web.

5.2.2 – The role of applications developed in WP2

On top of the above mentioned dissemination activities, an important form of dissemination, possibly the most effective, will consist in the deployment of the application described in WP2 and in their public accessibility. Beyond their use as test beds for the technologies developed within the consortium, and their use as sources of data for research purposes, the applications constitute a natural dissemination channel, because they are aimed at creating and supporting the activities of new social communities on the web.

5.2.3 - Contribution of Sony CSL

Sony CSL organises bi-annual public symposia in Paris on novel topics relevant to science, technology and society. During the duration of the project, there will be two symposia, in 2006 and 2008, respectively. The symposium of the year 2006 will be organised in form of an exhibition centered on the theme of intensive science, as an attempt to bridge art and science. In this context, the experimental tagging systems described in WP2, Task 2.1 will be shown. The work will also be disseminated in form of a catalogue of the symposium 2006.

Furthermore, Sony CSL seeks contact with the artistic community. To this end, CSL performs experiments on the image tagging and navigation system in collaboration with the photographer Armin Linke (http://www.arminlinke.com/). This work will be shown in exhibitions, such as the Moderna Museet in Stockholm. In addition, the experimental system will be tested by students of Armin Linke within workshops.

Task 5.3 Training activities and outreach

The training activities will include: (a) training activities for the researchers hired within the project, (b) organization of short courses aimed at the participants of the project and also at the participants of the other projects of the Complex System initiatives, (c) organization of tutorials (also web-based) and summer schools within the project workshops or within other initiatives, (d) coordination of the staff exchange initiative that will involve exchange of staff both among the partners of the consortium and with the partners of other projects. In addition, training activities are foreseen in the context of conferences or other kinds of gatherings, for example in the form of tutorials, contributions to summer schools, etc.

Since several activities of the present program are concerned with user-oriented applications and specifically web-based tools potentially involving very large user base, we expect a significant outreach. In fact a side effect of the implementation of this program will be the exposure of our user base to technologies developed by the Consortium and concepts related to Semiotic Dynamics.

Deliverables

D5.1 - (Task 5.1) Project presentation report (Month 4).
D5.2 - (Task 5.2) Web site for the project (Month 4).
D5.3 - (Task 5.3) A *White Paper* that will describe target problems and grand challenges for Semiotic Dynamics systems, clearly recognised by all and openly communicated to the scientific community (Month 11).
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).
D5.5 - (Task 5.3) Report on the impact, usability and user communities characterization of our web-based experiments and demos (Month 35).

Milestones and expected result

M5.1 - Identification of third parties (SMEs) suitable for the deployment of the web-based part of the dissemination plan (Month 11).
M5.2 - **M5.3** - The two Sony CSL bi-annual public symposia in Paris (2006 and 2008) will set important milestones to disseminate the scientific and technological results of the Consortium (Tentatively Months 5 and

29).

Workpackage 6 (WP6) - Management

Workpackage number	**6**	**Start date or starting event:**			**0**	
Workpackage title	**Management**					
Participant id		**1**	**2**	**3**	**4**	**5**
Person-months per participant:		**18**	**0**	**0**	**0**	**0**

Objectives

- 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 workpackage objectives, quickly reporting any problems to management.
- To provide evaluation reports to the Commission as required

Description of work

Task 6.1 Management

This workpackage will provide co-ordination of administrative and scientific work, including the arranging and recording of meetings. It will also provide the interface with the Commission in all matters, as well as publicity and information about the project, both to the media and globally through the World Wide Web.

Staff in this workpackage is responsible for the evaluation of the project in a continuous fashion through assessment against both project objectives and against the objectives of individual workpackages. They should monitor the reaching of milestones and the delivery of deliverables, and report to the management any problems that arise or are foreseen.

Deliverables

D6.1 - Provision of reports as required to the Commission.
D6.2-D6.4. - Yearly Management Report (see section B5) (Month1 11,23,35).

Milestones and expected result

M6.1 (Month 3) - Set up of the project information infrastructure (WWW pages, mailing list, ftp area etc.)
M6.2 to **M6.5** - Co-ordination and Management Meetings (months 0, 11, 23, 35).

8. Project resources and budget overview

In this section we will describe the resource needed to carry out the project and their disposition across the consortium. The planned duration of the project is 3 year.

Personnel

The principal resource required by the project is personnel. In Table 1 we indicate the contribution in person months for staff members of each partner and the person months for post-doc or phd students to be hired during the project. This table shows the critical mass that the consortium can bring to bear on the project. A part-time administrator (18 man-months during 3 years) is employed by the Coordinating partner (PHYS-SAPIENZA) to assist the coordinator in the administration of the consortium.

Partner	Cost Model	Staff	Hired Researchers
1. PHYS-SAPIENZA	FC	**28**	96
2. SONY-CSL	FC	**20**	60
3. UNI KO-LD	AC	3	46
4. UNIK	AC	10	45
5. UNI-SOTON	AC	10	36
Total		71	283

Table 1. Person months for each partner for the full duration of the project (bold is used for teams with the FC cost model).

Section 8.1 (below) describe the person-months per partner associated with each activity identified in the sections above (notice that in this table for partners adopting the AC Cost model we only include the man months for the people to be hired during the project).

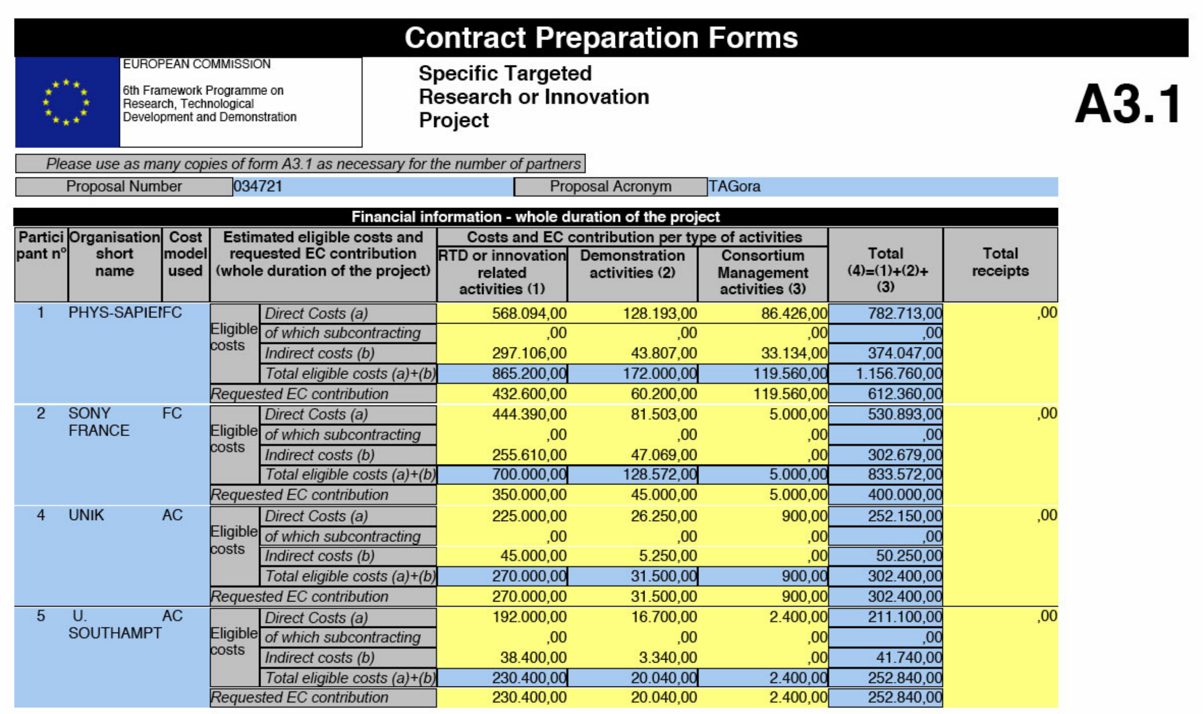
8.1 Efforts for the project (STREP/STIP Efforts Form in Appendix 1)

STREP Project Effort Form
Full duration of project

Project number (acronym): **TAGora**

	Partner 1 PHYS-SAPIENZA	Partner 2 SONY-CSL	Partner 3 UNI KO-LD	Partner 4 UNIK	Partner 5 UNI-SOTON	TOTAL PARTNERS
Research/innovation activities						
WP1 Emergent Metadata	15	9	3	3	16	46
WP2 Applications	0	20	13	15	0	48
WP3 Data analysis of emergent system properties	34	3	15	15	4	71
WP4 Modeling and simulations	60	32	12	8	12	124
Total research/innovation	98	64	43	41	32	280
Demonstration activities						
WP5 Dissemination and exploitation	15	16	3	4	4	42
Total demonstration	15	16	3	4	4	42
Consortium management activities						
WP6 Management	18	0	0	0	0	18
Total consortium management	18	0	0	0	0	18
Total per Partecipant	142	80	46	45	36	349
Overall TOTAL EFFORTS						349

8.2 Overall budget for the project (Forms A3.1 & A3.2 from CPFs)



Contract Preparation Forms

EUROPEAN COMMISSION

6th Framework Programme on
Research, Technological
Development and Demonstration

Specific Targeted
Research or Innovation
Project

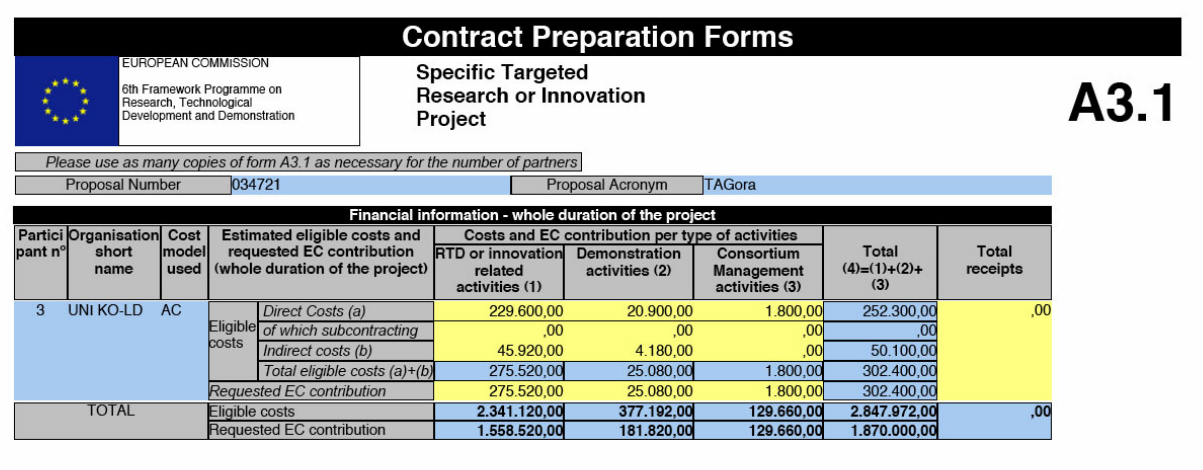
A3.1

Please use as many copies of form A3.1 as necessary for the number of partners

Proposal Number 034721 Proposal Acronym TAGora

Financial information - whole duration of the project

Partici pant n°	Organisation short name	Cost model used	Estimated eligible costs and requested EC contribution (whole duration of the project)		Costs and EC contribution per type of activities			Total (4)=(1)+(2)+ (3)	Total receipts
					RTD or innovation related activities (1)	Demonstration activities (2)	Consortium Management activities (3)		
1	PHYS-SAPIEI	FC	Eligible costs	*Direct Costs (a)*	568.094,00	128.193,00	86.426,00	782.713,00	,00
				of which subcontracting	,00	,00	,00	,00	
				Indirect costs (b)	297.106,00	43.807,00	33.134,00	374.047,00	
				Total eligible costs (a)+(b)	865.200,00	172.000,00	119.560,00	1.156.760,00	
			Requested EC contribution		432.600,00	60.200,00	119.560,00	612.360,00	
2	SONY FRANCE	FC	Eligible costs	*Direct Costs (a)*	444.390,00	81.503,00	5.000,00	530.893,00	,00
				of which subcontracting	,00	,00	,00	,00	
				Indirect costs (b)	255.610,00	47.069,00	,00	302.679,00	
				Total eligible costs (a)+(b)	700.000,00	128.572,00	5.000,00	833.572,00	
			Requested EC contribution		350.000,00	45.000,00	5.000,00	400.000,00	
4	UNIK	AC	Eligible costs	*Direct Costs (a)*	225.000,00	26.250,00	900,00	252.150,00	,00
				of which subcontracting	,00	,00	,00	,00	
				Indirect costs (b)	45.000,00	5.250,00	,00	50.250,00	
				Total eligible costs (a)+(b)	270.000,00	31.500,00	900,00	302.400,00	
			Requested EC contribution		270.000,00	31.500,00	900,00	302.400,00	
5	U. SOUTHAMPT	AC	Eligible costs	*Direct Costs (a)*	192.000,00	16.700,00	2.400,00	211.100,00	,00
				of which subcontracting	,00	,00	,00	,00	
				Indirect costs (b)	38.400,00	3.340,00	,00	41.740,00	
				Total eligible costs (a)+(b)	230.400,00	20.040,00	2.400,00	252.840,00	
			Requested EC contribution		230.400,00	20.040,00	2.400,00	252.840,00	



Contract Preparation Forms

EUROPEAN COMMISSION

6th Framework Programme on
Research, Technological
Development and Demonstration

Specific Targeted
Research or Innovation
Project

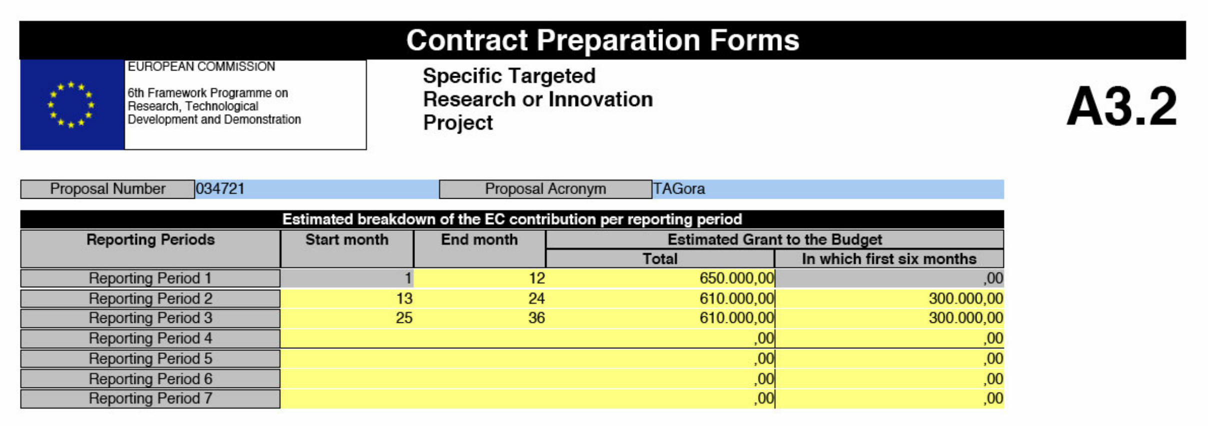
A3.1

Please use as many copies of form A3.1 as necessary for the number of partners

Proposal Number 034721 Proposal Acronym TAGora

Financial information - whole duration of the project

Partici pant n°	Organisation short name	Cost model used	Estimated eligible costs and requested EC contribution (whole duration of the project)		Costs and EC contribution per type of activities			Total (4)=(1)+(2)+ (3)	Total receipts
					RTD or innovation related activities (1)	Demonstration activities (2)	Consortium Management activities (3)		
3	UNI KO-LD	AC	Eligible costs	*Direct Costs (a)*	229.600,00	20.900,00	1.800,00	252.300,00	,00
				of which subcontracting	,00	,00	,00	,00	
				Indirect costs (b)	45.920,00	4.180,00	,00	50.100,00	
				Total eligible costs (a)+(b)	275.520,00	25.080,00	1.800,00	302.400,00	
			Requested EC contribution		275.520,00	25.080,00	1.800,00	302.400,00	
TOTAL			Eligible costs		**2.341.120,00**	**377.192,00**	**129.660,00**	**2.847.972,00**	**,00**
			Requested EC contribution		**1.558.520,00**	**181.820,00**	**129.660,00**	**1.870.000,00**	



Contract Preparation Forms

EUROPEAN COMMISSION

6th Framework Programme on Research, Technological Development and Demonstration

Specific Targeted Research or Innovation Project

A3.2

Proposal Number	034721	Proposal Acronym	TAGora

Estimated breakdown of the EC contribution per reporting period

Reporting Periods	Start month	End month	Estimated Grant to the Budget	
			Total	In which first six months
Reporting Period 1	1	12	650.000,00	,00
Reporting Period 2	13	24	610.000,00	300.000,00
Reporting Period 3	25	36	610.000,00	300.000,00
Reporting Period 4			,00	,00
Reporting Period 5			,00	,00
Reporting Period 6			,00	,00
Reporting Period 7			,00	,00

8.3 Management level description of resources and budget

In this section we will describe the resource needed to carry out the project and their disposition across the consortium.

Travel, Conferences, and staff exchanges

To favour the project integration, dissemination and training activities we foresee the following activities:

- Project meetings. Founds are allocated so to cover 6 meetings.
- Dissemination: Participation to international conferences, workshops, summer schools.
- Staff exchanges. Founds are allocated so to allow EU partner to exchange staff with all the other partners of the TAGora project and with the partners of the other project of the Complex System Initiative.
- Conferences of the Complex System Initiative. Founds are allocated to organize an open and one closed conferences of the Complex System Initiative.
- A budget, provisionally allocated to PHYS-SAPIENZA, will be used for dissemination activities (e.g. inviting journalists to the bi-annual conferences, commissioning articles, organizing industry-events etc.).
- A budget, provisionally allocated to PHYS-SAPIENZA, is foreseen to contribute to the activities of the 'complex systems' cluster.

The total cost of these activities is € 248.150 (+ overhead). The exact allocation of these founds per partner is described below.

Equipment

The total cost for hardware and computer equipments is € 78.610 (+ overhead). In particular PHYS-SAPIENZA plan of buying a small Linux cluster for data analysis and high-performance simulations.

Consortium Management

Apart from the cost of a part-time administrator assisting the coordinator of the project, founds for consortium management include:

- certification of cost statements
- consumables costs
- supplementary meeting venue costs and travel/subsistence costs of the consortium administrator

Overall, coordination costs will amount to € 96.526 (+ overhead).

Detailed description of Resource Needed to Carry Out the Project

Phys-Sapienza will support the project with 1 full professor, 1 associate professor, 1 senior researcher (28 man months in total) and will hire three postdoctoral researchers (96 man months in total). Moreover it will hire a part-time administrator (for 18 man-months) to assist the coordinator in the management of the Consortium. Total cost for staff and researchers to be hired: EUR 609.597. Total cost for the project management, project meetings, dissemination, and staff exchange: EUR 143.113. This amount will be used also to invite researchers outside the Complex System Initiative but with related expertise and in particular: Martin Nowak, William S.-Y. Wang, Clay Shirky. Cost for new computer equipment is EUR 30000. Cost for audit certificates: EUR 6.000. Cost for overhead (61.36% of the personnel total cost): 374.050 Total cost: EUR 1.156.760. Requested contribution: EUR 612.360. Key personnel assigned: Vittorio Loreto (coordinator), Luciano Pietronero, Ciro Cattuto, Andrea Baronchelli.

Sony CSL will support the project with 1 part-time professor, 1 part-time researcher (20 man-months), 1 full-time associate researcher for 24 months (24 man-months) and 1 full-time associate researcher for 36 months (36 man-months). Total effort is 80 man-months. Total cost for staff and researchers to be hired: EUR 490.621. Cost for overhead (62% of personnel total cost): EUR 302.679. Total cost for project meetings, dissemination, and staff exchange are: EUR 29.272. Cost for new computer equipment is EUR 6.000. Cost for audit certificates: EUR 5.000. Total cost: EUR 833.572. Requested contribution: EUR 400.000. Key personnel assigned: Luc Steels, Peter Hanappe, Melanie Aurnhammer and one person to be recruited.

Uni Ko-Ld will support the project with 1 professor and two junior researchers (46 person months). Total cost for staff is: EUR 18.000 (3 man-months). Total cost for researchers to be hired: EUR 220.800. Total cost for project meetings, dissemination, and staff exchange are: EUR 22,000. Cost for new computer equipment is EUR 7700. Cost for audit certificates: EUR 1800. Cost for overhead (20% of the total cost without audit): 50,100 Total cost: EUR 302,400. Requested contribution: EUR 302,400. Key personnel assigned: Steffen Staab (coordinator), Olaf Görlitz, Klaas Dellschaft.

UniK will support the project with 1 full professor, one senior and three junior researchers (10 person months in total, cost: EUR 60.400 and will hire two researchers (total 45 person months). Total cost for researchers to be hired: EUR 216.000. Total cost for project meetings, dissemination, and staff exchange are: EUR 23.000. Cost for new computer equipment is EUR 12.250. Cost for audit certificates: EUR 900. Cost for overhead (20% of the total cost without audit): 50.250 Total cost: EUR 302.400. Requested contribution: EUR 302.400. Key personnel assigned: Gerd Stumme (coordinator), Andreas Hotho, Christoph Schmitz.

Uni-Soton will support the project with 1 professor and 2 senior researchers (12 person months) and will hire 1 postdoctoral researcher (36 person months). Total cost for staff is: EUR 145.042 (10 man-months). Total cost for a researcher to be hired: EUR 154.909. Total cost for project meetings, dissemination, and staff exchange are: EUR 31.200. Cost for new computer equipment is EUR 22.665. Cost for audit certificates: EUR 2.400. Cost for overhead (20% of the total cost without audits): 41.666 Total cost: EUR 252.840. Requested contribution: EUR 252.840. Key personnel assigned: Prof Nigel Shadbolt (coordinator), Dr Harith Alani, Dr Kieron O'Hara.

8.4 Project clustering and collaboration within the FET CS cluster

The project will foresee resources to actively participate in the FET proactive initiative "Complex Systems" (CS). The projects of this cluster consist of all Integrated Projects funded as part of the first call in CS – EVEGROW, DELIS, PACE, ECAGENT, all projects funded in the second call (including this project) – COAST, PERPLEXUS, EMIL, TAGORA, EURACE, GENNETEC – and projects funded form the FET OPEN – NEW TIES, BISON, ISCOM (see www.cordis.lu./fet/co.htm).
The coordination action ONCE-CS (www.once-cs.net) was set up to facilitate interaction between projects. Information pertinent to the cluster will be available on the website of ONCE-CS.
Budget to contribute to the functioning of the cluster will be set aside (specific items are mentioned below) by the project. The project will also propose events and activities specific to the project that could be of general interest to the CS cluster (eg summer schools or workshops of general interest).

The minimum contributions made by the project include:

- Actively contribute to the yearly conference in CS (the 3rd conference ECCS 2006 will be held in Oxford in September). To this end budget will be set aside to send researchers to this conference and to finance WS during this conference (approximately 5000euros per year to cover travel and registration fees for a representative group of researchers from the project, invited speakers to the conference in the area of the project, and to organise a workshop during the conference).

- Workshops are planned to link research in CS to industry and policy making. The project foresees budget to contribute to the organisation of these workshops (one is planned for 2006 and the project sets aside 7500Euros for this event). A further workshop is foreseen for 2007. These WS should help to disseminate the idea of CS thinking to a selected group of industrialists and decision makers.

- The project will contribute to the editing and maintenance of the roadmap. This includes timely input to the content of the roadmap (edited by ONCE-CS) and budget to send representatives to think-tank meetings organised by ONCE-CS.

- Actively promote the work of the CS cluster, disseminate information internally and externally through several media (web sites, Newsletters, reports, presentations, press releases, articles in the specialized and popular press, etc.).

- Provide yearly up-dated presentation material highlighting project results, in the form of publications, slides, pictures, videos, press releases (for web sites, reports, newsletters, etc.).

- Contribute (logistically and financially) to the organization of yearly cluster reviews.

9. Ethical issues

Ethical issues are obviously of enormous importance in the context of European research projects, particularly for projects involving future and emergent technologies for which the application domains are not yet very clear and hence the public or legislative bodies are not yet in a position to formulate clear policy. The project partners have done an extensive investigation in what kind of ethical issues may arise and how they could address the various ethical issues that may come up. Although no specific work-packages have been foreseen, we plan to infuse the total project with ethical considerations. Failure to handle the ethical issues would obviously be irresponsible from the viewpoint of scientists involved in this project, even if their work involves mostly mathematical models and computer simulations. The scientific committee will spend a considerable effort to ensure that ethical issues are fully taken into account in all aspects of the project.

Full investigation of all ethical issues raised by this project, and particularly the many legal ramifications in complex national and international regulations, obviously has to wait until the project is operational, partly because of the huge complexity involved and the need to be knowledgeable about the regulatory process inside the community and its various member states. Legal advice may be necessary and the project hopes to rely on the know how of the project officers with respect to ethical issues or EC-policy related issues.

All activities in the project will be continuously examined from the viewpoint of ethical issues and from the viewpoint of national and international regulations that might be applicable. Corrective actions will be taken as soon as there is a risk of unethical behavior or if the activities in the project do not fit with the national and international regulatory requirements.

More concretely,

- For all experiments that involve human subjects, strict regulations will be followed respecting the rights of human subjects in such experiments. These regulations depend on the location where the experiments will be carried out.
- For the purpose of data analysis, the partners will only handle anonymous data or data in which the user cannot be identified.
- Computing in ubiquitous environments may raise various privacy issues. The project is therefore committed to respecting the legislation with respect to privacy of information of citizens. For example, no databases with personal materials collected during the project will be sold under any pretext. They will remain private and will be destroyed when the scientific results have been achieved
- All scientific papers will be written and published with the strictest ethical criteria, with respect to plagiarism, honesty of reference, etc. No work will be published that is not originating with the authors of the publication.
- Application of industrial results from other projects will take into account the full intellectual property rights of their owners.
- All mechanisms developed in the project are targeted to ensure that there is an open access to the benefits of the information society for all. This is why the project promotes open standards and why all results are made as much public as possible.

The project is committed to pursue a very strong anti-discriminatory policy in the sense that research results from the project will be communicated openly to any member of the public that requests it and there is no limit put on access to the result of the project. Also in terms of hiring there is an anti-discriminatory policy, in particular with respect to woman scientists (see gender issues).

Ethical issues checklist

Table A. Proposers are requested to fill in the following table

Does your proposed research raise sensitive ethical questions related to:	**YES**	**NO**
Human beings		X
Human biological samples		X
Personal data (whether identified by name or not)	X	
Genetic information		X
Animals		X

Please indicate whether the proposal involves	YES	NO	UNCERTAIN
· Research on human beings			
Persons not able to give consent		X	
Children		X	
Adult healthy volunteers		X	
· Human biological samples			
Human foetal tissue/cells		X	
Human embryonic stem cells		X	
· Human embryos		X	
· Human genetic information		X	
· Other personal data			
Sensitive data about health, sexual lifestyle, ethnicity, political opinion, religious or philosophical conviction		X	
· Animals (any species)			
Non- human primates		X	
Transgenic small laboratory animals		X	
Transgenic farm animals		X	
Cloning of farm animals		X	
· Research involving developing countries (e.g. clinical trials, use of human and animal genetic resources…)		X	

· Dual use		X	

Table B. Proposers are requested to confirm that the proposed research does not involve:

- Research activity aimed at human cloning for reproductive purposes,
- Research activity intended to modify the genetic heritage of human beings which could make such changes heritable[12]
- Research activity intended to create human embryos solely for the purpose of research or for the purpose of stem cell procurement, including by means of somatic cell nuclear transfer.

	YES	**NO**
Confirmation : the proposed research involves none of the issues listed in Table B		X

10. Other issues

Gender issues

The project is within a scientific domain where there is so far clearly less participation from female researchers. This is seen in terms of the lack of publications and conference participation in this field by female researchers. The project cannot correct a general situation where there are not enough female students in the areas covered by the project (particularly for computer science). In this respect they feel strongly that national governments and the EU IT program should make resources available to educational institutions to encourage intake of female students and to fund special programs to motivate female students.

Nevertheless the project will launch a series of actions embedded in the project which constitute a Gender Action Plan.

Action 1. We propose that the EU project officer monitoring this project, as well as the reviewers of the project's progress, should be female specialists.

Action 2. We recommend to all partners that if new personnel is being hired to carry out the research discussed in this project, there is specific emphasis on hiring female researchers.

Action 3. The project foresees a number of dissemination actions in the form of schools and workshops. A special action will be undertaken to ensure that female participation in these dissemination activities is high. This will be done by using communication channels of female scientists for announcing the events, by biasing the selection of candidates, and by seeking female lecturers as parts of these dissemination activities.

[12] Research relating to cancer treatment of the gonads can be financed

Action 4. In all communications of the project special care will be taken to use gender-neuter language.

Appendix A - Consortium description

A.1 Participants and consortium

From the description of research objectives it should be clear that this is a project that requires a wide variety of skills and competences including Statistical Physics, Complex System Theory, Information Technology, Robotics, Linguistics.

The chosen Consortium consists of 5 teams that:

- provide all the requested competences;
- have different profiles: 4 partners are public institutions involved in research and higher education, 1 (SONY-CSL) is an industrial partner;
- include leading scientists in the corresponding fields.

1. Physics Department, "La Sapienza" University, Roma, Italy (Phys-Sapienza). Phys-Sapienza is one of the largest physics department in Italy. It includes more than 250 scientists, among professors and researchers, and more than 66 research groups with topics ranging from high-energy physics to condensed matter theory, statistical mechanics and astrophysics. The Phys-Sapienza team participating to the consortium is composed by a world-level research group in the whole area of statistical physics, information theory and complex systems. Traditionally the activity of this group has been focused on several main topics: frustrated and glassy systems, granular media, fractal growth, dynamical systems, superconductivity. In addition in the last few years several new projects have been launched among which:

- agent-based modeling in linguistics;
- information theory applied to time-series analysis, linguistics and genomics,
- theory of complex networks in technological, social and biological or bio-inspired systems;
- assimilation dynamics in social systems.

Phys-Sapienza team brings into the consortium its research experience on: (a) developing new theoretical tools to collect and analyse data; (b) 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; (c) constructing theoretical approaches which could provide with different levels of abstraction and a feedback for new experiments and studies.
Phys-Sapienza team is involved in the context of the EU project ECAgents (IST-1940). This project focuses on the development of a new generation of embodied agents that are able to interact directly (i.e., without human intervention) with the physical world and to communicate between them and with other agents (including humans). See the details below in the description of the SONY team. Research will be lead by Prof. Vittorio Loreto (http://pil.phys.uniroma1.it/~loreto/). Principal researchers will include : Prof. Luciano Pietronero (http://pil.phys.uniroma1.it/~luciano/), Dr. Andrea Baronchelli. An additional resource will be represented by Dr. Ciro Cattuto, who holds a research “New talent” grant from the Museo Storico della Fisica e Centro Studi e Ricerche “Enrico Fermi”

(http://www.centrofermi.it), to pursue investigations on "Communication and self-organization in biological, technological and social systems".

REFERENCES
- D. Benedetto, E. Caglioti and V. Loreto, "Language Trees and Zipping", *Phys. Rev. Lett*. **88** 048702 (2002).
- G. D'Anna, P. Mayor, A. Barrat, V. Loreto and F. Nori, "Observing Brownian motion in vibration-fluidized granular matter", *Nature*, **424**, 909 (2003).
- F. Radicchi, C. Castellano, F. Cecconi, V. Loreto and D. Parisi, "Defining and identifying communities in networks", *PNAS*, **101**, 2658 (2004).
- A. Baronchelli, M. Felici, E. Caglioti, V. Loreto and L. Steels, "Sharp transition towards shared vocabularies in Multi-Agents Systems", physics/0609075 at xxx.lanl.gov, submitted to PNAS.

2. Sony Computer Science Laboratory, France. Sony CSL Paris was founded in 1996 and is a small but booming research cell, focussing on four areas: personal music experience, computational neuroscience, developmental cognitive robots, and self-organising communicating systems.

Research in Personal Music Experience focuses on the future of musical listening by building prototypes of interactive devices and ethnographic experiments to see what people find exciting in music and how new ways of listening integrate into their lives. The Computational Neuroscience group uses mathematical and computational techniques to make realistic models of the brain, in particular the cerebellum. This is expected to yield radically new ideas for building adaptive machines with life-like learning behavior. The Developmental Cognitive Robotics group tries to work out a scenario in which an autonomous embodied robot in interaction with the environment, other robots, and human beings, can bootstrap cognitive behavior and intelligence. Research in self-organising communication systems investigates through computational simulations and mathematical models how a group of autonomous agents could be able to invent and negotiate a communication system similar to human natural languages.

Sony-CSL has been working on various topics that are directly linked to the themes of this research project in collaboration with other entities in Sony Europe that are closer to applications. The most important work so far has focused on researching new ways to access large music databases that could support emergent semantics as well as on Semiotic Dynamics. In particular, the EU project "Semantic HIFI" (IST-507913) can provide insights in experimental sharing and interaction systems in the domain of music. In the context of large scale digital music, the goal of this project is to develop a new generation of HIFI systems, offering new functionality for browsing, interacting, rendering, personalizing and editing musical material. This next generation of Hard-disk based HIFI systems will change drastically the relation of home users to music and multimedia content. Users will be able to interact with music up to the point of blurring the traditional limits between playing, performing and remixing (Pachet, F., 2004). These HIFI systems will be listening stations as much as open instruments. Systematic use of metadata extraction and exploitation techniques will allow semantic or thematic browsing in large content catalogues over Web and file sharing systems. Converging with the TV, DVD, Game Station and 5.1 set-ups, HIFI systems will also bring 3D audio real-time navigation in the sound scene. Results from the Semantic HIFI project are expected to yield important information on the way user interact with music data.

In the domain of Semiotic Dynamics, CSL researchers have pioneered algorithms for studying how populations of agents are able to arrive at a system of shared tags (a lexicon). Characteristic for our approach is that we also model the process by which categories emerge and get coordinated through tagging (Language games reference). This is ongoing research at CSL which is also investigated in the context of the EU project ECAgents (IST-1940). This project focuses on the development of a new generation of embodied agents that are able to interact directly (i.e., without human intervention) with the physical world and to communicate between them and with other agents (including humans). The specific aims of the project are: (1) designing a new generation of embodied agents able to evolve autonomously, self-organize, and operate reliably in a dynamic environment, (2) setting up the conditions that allow a population of embodied agents to develop a shared communication language and to share knowledge, and (3) identifying new methods and algorithms that allow to engineer systems able to self-organize and to display properties emerging from the interactions between themselves and with the external environment. This research will be the foundation for making reasonably realistic agent-based simulation models of tagging in populations of agents (WP4).

REFERENCES

- Steels, L. and P. Hanappe (submitted) Interoperability through Emergent Semantics. A Semiotic Dynamics Approach. Submitted to Journal on Data Semantics
- Steels, L. Evolving grounded communication for robots. Trends in Cognitive Science, 7(7):308-312 July 2003.
- Pachet, F. The HiFi of the Future: Toward new modes of Music-ing. In Perrot, X., editor, Proceedings of ICHIM 04, 2004.
- Zils, A. and Pachet, F. Automatic Extraction of Music Descriptors from Acoustic Signals using EDS. Proceedings of the 116th AES Convention, May 2004.

3. Institute for Computer Science, Faculty of Computer Science of the University of Koblenz Landau. Universitaet Koblenz-Landau (D) has four departments, one of them being the department for computer science; there are currently about 11,000 students subscribed at the university, about 1300 of whom study in the department of computer science. The computer science department comprises 20 professors. It has a track record of teaching and research in informatics for over 25 years offering studies in the areas of computer science, computer visualistics, business informatics and information management. The institute for informatics includes 8 professors (Furbach, Beckert, Ebert, Rosendahl, Steigner, Zöbel, Staab, Lautenbach) teaching and researching in the core areas of computer science. In particular, there is intensive research in the areas of Semantic Web, knowledge-based systems and reasoning (Prof.s Beckert, Furbach, Staab). The research group on information systems and the semantic web (ISWeb) has been founded by Prof. Staab in Fall 2004, it includes 10 researchers and is well-known in the Semantic Web area. Research will be lead by Prof. Steffen Staab (http://www.uni-koblenz.de/~staab/). Prof. Dr. Steffen Staab is professor for databases and information systems in the institute for informatics of the University of Koblenz-Landau. He heads the research group on information systems and the semantic web (ISWeb), which is participating in several EU IST integrated project on semantic multimedia, metadata management and ontology management. Before his current position Prof. Staab studied at the University of Erlangen-Nuremberg, University of Pennsylvania, and University of Freiburg, where he received his prediploma, his M.S.E. and his Dr. rer. nat., respectively. He then did consulting at the Fraunhofer institute for industrial engineering (IAO) in Stuttgart and holds a position as project manager and lecturer at the University of Karlsruhe. Prof. Staab has a wide range of research interests including semantic web, text mining, ontologies, peer-to-peer and service management with semantic descriptions, which led to over 100 refereed publications and 7 books, including the recent Handbook on Ontologies. Recently, he coordinated the 5FP IST project "SWAP - Semantic Web and Peer-to-peer", which culminated into a new book of the same name that deals with the life cycle of metadata in dynamic peer-to-peer systems. He is on the editorial board of IEEE Intelligent Systems, Journal of Web Semantics, Int. J. on Human-Computer Studies and Information Technology & Tourism. Prof. Staab is co-chair of the European Conference on Knowledge Engineering and Knowledge Management 2006 and tutorial chair for the ACM Intelligent User Interfaces conference 2006. Principal researchers will include: Dr. Olaf Görlitz.

REFERENCES

- S. Handschuh, S. Staab (eds.). *Annotation for the Semantic Web*. IOS Press, June 2003.ù
- S. Staab, H. Stuckenschmidt (eds.). *Semantic Web and Peer-to-Peer*. Springer, Fall 2005.
- S. Staab, R. Studer (eds.). *Handbook on Ontologies.* International Handbooks on Information Systems, Springer Verlag, 2004.

4. Hertie Chair of Knowledge & Data Engineering, University of Kassel The research unit Knowledge & Data Engineering in the department of Mathematics/Computer Science 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. Work at L3S focuses on innovative information systems, on learning and knowledge technologies and on innovative concepts and infrastructures for training and continuing education in academia and industry. L3S projects include research, consulting, and technology transfer in all of these areas, provision of infrastructure and support for innovative teaching and learning technologies at the participating universities, and collaboration with both German and international standardization bodies. Today, over 50 researchers work at the L3S in the areas of Semantic Web and Digital Libraries, Peer-to-peer Information Systems, eLearning, Industrial Informatics, and Mobile / Distributed Computing and Networks. A sixth area (Grid Computing) is currently being started. Via L3S, the research unit participates in the NoE Prolearn. Based on its roots at the Institute AIFB at the University of Karlsruhe, the research unit has extensive experience in managing European and national projects. The research team will be directed by Prof. Dr. Gerd Stumme. Gerd Stumme is Full Professor of Computer Science. He is leading the research group on Knowledge and Data Engineering at the University of Kassel, and full member of the Research Center L3S. He earned his PhD in 1997 at Darmstadt University of Technology, and his Habilitation at the Institute AIFB of the University of Karlsruhe in 2002. In 1999/2000 he was Visiting Professor at the University of Clermont-Ferrand, France, and Substitute Professor for Machine Learning and Knowledge Discovery at the University of Magdeburg in 2003. Gerd Stumme published over 80 articles at national and international conferences and in journals, and chaired several workshops and conferences. He is member in the Editorial Boards of the Intl. Journal on Data Warehousing and Mining and of the International Conference on Conceptual Structures, and was also member of several conference and workshop Program Committees. Gerd Stumme led several national and European projects, eg, the current national project "Personalized Access to Distributed Learning Resources (PADLR)". Additionally he is co-chairing the Web Mining Forum of the European Network of Excellence KDNet. Dr. Andreas Hotho will take also part to the project. Andreas Hotho holds a Ph.D. from the University of Karlsruhe, where he worked from 1999 to 2004 at the Institute of Applied Informatics and Formal Description Methods (AIFB) in the areas of text, data, and web mining, semantic web and information retrieval. He earned his Master's Degree in information systems from the University of Braunschweig (Germany) in 1998. Since 2004 he is a senior researcher at the University of Kassel. His focus is on the combination of machine learning/data mining and semantic web, called semantic web mining, and especially on text clustering/classification with background knowledge. He was involved in organizing several workshops in conjunction with ECML/PKDD and KDD conferences with topics rela

REFERENCES

- P. Cimiano, A. Hotho, G. Stumme, J. Tane: Conceptual Knowledge Processing with Formal Concept Analysis and ontologies. In Proc. 2nd International Conference on Formal Concept Analysis (ICFCA'04), Springer, Heidelberg 2004 (Invited Talk)
- F. Dau, M.-L. Mugnier, G. Stumme (eds.): Conceptual Structures: Common Semantics for Sharing Knowledge. Proc. 13th. Intl. Conf. on Conceptual Structures. LNAI 3596. Springer, Heidelberg 2005
- B. Ganter, G. Stumme, R. Wille (eds.): Formal Concept Analysis - Foundations and Applications. State of the Art, LNAI, Springer 2005.
- A. Hotho, S. Staab, G. Stumme: Explaining Text Clustering Results using Semantic Structures. In: N. Lavrac, D. Gamberger, L. Todorovski, H. Blockeel (Hrsg.): Knowledge Discovery in Databases: PKDD 2003. LNAI 2838. Springer, Heidelberg, 2003, 217-228
- G. Stumme: Off to New Shores - Conceptual Knowledge Discovery and Processing. Intl. J. Human{Comuter Studies (IJHCS) 59(3), September 2003, 287-325ted to semantic web mining.

5. Department of Electronics and Computer Science University of Southampton The School of Electronics and Computer Science at Southampton is a world-leading centre of excellence for research, teaching, enterprise and innovation and is now one of the largest schools of its kind in the world. The School consistently achieves the highest ratings and was rated 5* by the RAE in 2001, and is now rated 6* by HEFCE. Within the school, the IAM (Intelligence, Agents, Multimedia) Group focuses on the design and application of computing systems for complex information and knowledge processing tasks. With around 120 researchers, we are international leaders in the three major themes that converge in the Group's tripartite title; Intelligence: Examining the fundamental principles of intelligent and adaptive behavior and developing methods and services for acquiring, modeling, reusing, retrieving, publishing and maintaining knowledge; Agents: Devising new methods and models for inter-agent interactions such as cooperation, coordination, auctions and negotiation, new mechanisms for establishing trust and reputation in open systems, and pioneering work on agent-oriented software engineering; Multimedia: Investigating the basic principles and applications of multi-modal communication, hypermedia and document management in large scale open systems such as digital libraries and the Semantic Web, and developing context aware, personalised information management systems. These three research themes also combine synergistically in a number of grand challenges for computer science - including grid computing, peer-to-peer systems, sensor networks, the semantic web, and pervasive computing environments. One particular project that our TAGora team is heavily involved in is a multi million pounds EPSRC IRC in Advanced Knowledge Technologies (AKT). AKT aims to develop and extend a range of technologies providing integrated methods and services for the capture, modelling, publishing, reuse and management of knowledge. A number of AKT technologies will be reused and perhaps further developed in TAGora. One example is CSAktive Space, which won the Semantic Web challenge in 2003. CSAktive Space contains, and inter-relates, automatically-harvested and updated information from various academic sources, and presents the results to users in digested and knowledge-enabled ways. Another relevant AKT application is OntoCoPI, which attempts to uncover Communities of Practice by applying a set of ontology-driven network analysis techniques that examine the connectivity of instances in the knowledge base with respect to the type, density, weight, and time of these connections. Key researchers in Southampton: Nigel Shadbolt is a professor of artificial intelligence. He led a consortium of five Universities that secured an EPSRC Interdisciplinary Research Collaboration in Advanced Knowledge Technologies (AKT). He is the Director of this multi-million pound, six-year research programme. He is Editor in Chief of IEEE Intelligent Systems, an Associate Editor of the International Journal for Human Computer Systems and on the editorial board of the Knowledge Engineering Review and the Computer Journal. He is a member of various national committees including the UK e-Science Technical Advisory Committee (TAG) and the UK EPSRC Strategic Advisory Team (SAT) for ICT, and is currently a Vice President of the British Computer Society. Kieron O'Hara is a senior research fellow in Electronics and Computer Science at the University of Southampton, working on the Advanced Knowledge Technologies project, on which he is a co-PI. He researches in the politics and epistemology of technology, and is the author of five books in the field. His particular focus is on the Semantic Web, and its implications for social interaction. He is a member of the EPSRC Memories for Life network, has worked on a number of the UK Department of Trade and Industry's Foresight programmes, and is currently editing a special issue of the International Journal of Human-Computer Studies on ontologies and knowledge representation. Harith Alani is a Research Fellow with the Intelligence, Agents, Multimedia Group in the School of Electronics and Computer Science, affiliated with the Advanced Knowledge Technologies (AKT) IRC. His current research activities are in semantic representation and analysis of communities of practice, ontology change management, and ontology ranking and minimisation. He has been a co-organiser of a number of knowledge management related workshops in various international conferences.

REFERENCES

- Alani, H., S. Dasmahapatra, K. O'Hara, and N. Shadbolt. (2003). "Identifying Communities of Practice through Ontology Network Analysis." IEEE Intelligent Systems 18(2): 18-25.
- Alani, H., S. Kim, D.E. Millard, M.J. Weal, Wendy Hall, P.H. Lewis, and N. Shadbolt (2003). "Automatic Ontology-based Knowledge Extraction and Tailored Biography Generation from the Web." IEEE Intelligent Systems 18(1): 14-21.
- Alani, Harith and Gibbins, Nicholas and Glaser, Hugh and Harris, Stephen and Shadbolt, Nigel (2005) Monitoring Research Collaborations Using Semantic Web Technologies. In Proceedings 2nd European Semantic Web Conference (ESWC), Crete.
- Glaser, H., Alani, H., Carr, L., Chapman, S., Ciravegna, F., Dingli, A., Gibbins, N., Harris, S., schraefel, m. c. and Shadbolt, N. (2004) CS AKTive Space: Building a Semantic Web Application, in Bussler, C., Davies, J., Fensel, D. and Studer, R., Eds. The Semantic Web: Research and Applications (First European Web Symposium, ESWS 2004), pages pp. 417-432. Springer Verlag.

Complementarity between participants

As an example of complementarity between participants it should be stressed that the PHYS-SAPIENZA team and the SONY-CSL team are strongly interacting already in the framework of the project ECAgents (IST-1940). The two teams are working in close collaboration to bridge the gap between the Statistical Physics and Complex systems communities (PHYS-SAPIENZA) on the one hand and the Computer Science and Linguistics communities (SONY-CSL) on the other hand. Several common papers are coming out as well as several common initiatives: Vittorio Loreto and Luc Steels are co-organizing a workshop on Semiotic Dynamics in November 2005 and a general conference on “Semiotic Dynamics, Language and Complexity” in the framework of the School of Complexity of the CCSEM, Erice December 2005.

In this perspective the SONY-CSL, beyond its status of industrial partner, represents the ideal link between the complex system community (here represented by PHYS-SAPIENZA) and the computer science/IT world. SONY-team has already collaborated with UNI KO-LD and UNIK and those teams are already closely linked with the UK UNI-SOTON team. For instance they are co-organizing the “Semantic Network Analysis” workshop at the ISWC2005 (International Semantic Web conference), Galway, Ireland November 2005. The presentation of this proposal has already represented an excellent opportunity to merge all the different communities and make their interests to focus on a common and challenging project.