Grant Agreement number: 288024
Project acronym: LiMoSINE
Project title: Linguistically Motivated Semantic aggregation engiNes
Funding Scheme: Collaborative project
Date of latest version of Annex I against which the assessment will be made:

Periodic report: 1st X 2nd □ 3rd □ 4th □
Period covered: from 1 November 2011 to 31 October 2012

Name, title and organisation of the scientific representative of the project's coordinator:
Prof. dr. M. de Rijke, University of Amsterdam
Tel: +31 20 525 76 93
Fax:
E-mail: m.derijke@uva.nl

Project website address:
http://www.limosine-project.eu
1. Publishable summary

1. Goal and challenges
We increasingly live our life online. Information is accumulated on a wide range of human activities, from science and facts, to personal content, opinions, and trends. Across the globe, people's knowledge, experiences and interactions effortlessly find their way to online outlets, alongside traditional edited content, ready to be shared with millions.

LiMoSINe will integrate the research activities of leading researchers across diverse topics with a view to enabling new kinds of language-based search technology. The LiMoSINe vision is to transition access to online information from a document-centric search paradigm focused on returning disconnected atomic pieces to a truly semantic aggregation paradigm. In this new paradigm, machines will understand a user's intent, discover and organize facts, identify opinions, experiences and trends, all from inherently multilingual online sources and open knowledge repositories. LiMoSINe's aggregation engines will automatically organize search results in semantically meaningful ways.

LiMoSINe has the following objectives:
- To enable semantically structured access to multi-lingual online content;
- To integrate deep linguistic processing in information extraction;
- To support semantic mining where data-driven patterns are made human interpretable using the web of data;
- To develop evaluation methods for rigorously assessing the effectiveness of semantic search and semantic aggregation in a lab-based setting;
- To exploit its research results in three demanding multilingual use cases:
  i. open-domain community question answering,
  ii. online reputation management in a professional task-based setting, and
  iii. intelligent content annotation and search on a photo-sharing platform.

The components of LiMoSINe will be integrated through web services with solutions currently in place at the project's use case owners.

2. Summary of the activities carried out

Semantic scenario refinement (WP2)
In order to lay the foundation for the research activities related to Information Access over Social Media, the team started its work by creating an overview of the state of the art approaches and tools. The main challenges associated to document annotation, mining, search and recommendation have been identified, focusing on the three application scenarios considered in the LiMoSINe project: Community Question Answering, Multimedia Tagging, and Online Reputation Management (ORM).

The novelty of the tasks to be undertaken in many cases meant the lack of appropriate datasets for research and evaluation. Therefore, data harvesting became one of the main priorities, producing at the end of the first year a number of valuable data collections:
• Community Question Answering corpus based on a large sample from Answerbag QA collection;
• Image Annotation corpus that comprises over 40,000 images extracted from Flickr and enriched with metadata;
• Several ORM corpora that embrace two manually annotated Twitter datasets created for the RepLab evaluation campaign and two manually annotated corpora built for studying the problem of determining opinion targets.

The first evaluation campaign for ORM systems, CLEF RepLab, took place,\(^1\) coordinated by LiMoSINe partners UNED, UvA and Llorente & Cuenca. The organisation of this benchmarking activity implied, in the first place, a thorough task definition. A special effort has been dedicated to setting up the evaluation methodology and studying suitable evaluation metrics for the problems defined. RepLab allowed us to bring together information access researchers and representatives from the ORM industry, inspiring a lot of interest, judging by the number of groups registered for participation (39).

One of the main outcomes of RepLab is the set of created test collections that enable systematic comparison of algorithms and reliable benchmarking of commercial ORM systems. The RepLab gold standard, which contains manual annotations made by reputation management experts, constitutes a useful source not only for evaluating systems, but also for obtaining a better understanding of the fundamental notions implied by reputation management.\(^2\)

In the future we expect to refine the chosen evaluation methodology and achieve more efficient algorithms for Community Question Answering, Multimedia Tagging and Online Reputation Management. Another important goal is to define and collate a common dataset appropriate for all the semantic aggregation scenarios and reusable by all the partners involved in the LiMoSINe project.

**Information extraction through deep linguistic analysis (WP3)**

The growing demand of online users for new functionalities of search engines, which is going far beyond simple keyword-based search, brings us deeply into the natural language processing territory. There is a growing need for a deeper semantic interpretation of data, requesting a semantic aggregation model. At the core of such a linguistically motivated semantic aggregation model is a representation that is able to capture various objects of interest (e.g., entities, opinions) and their relationships. One of the goals of the project is to develop such a semantic model extractor, which is an automatically extracted semantic representation of various levels of linguistic annotation. The semantic model is based on disambiguated entities, relations between them, subjective expressions, opinion holders and relations among these pieces of semantic information.

During the first year of LiMoSINe, an initial prototype of the semantic model extractor has been developed using the Apache UIMA (Unstructured Information Management Architecture) framework and deployed as web service. An overview of the overall pipeline of the semantic model extractor is shown in Figure 1.

---

\(^1\) [http://www.limosine-project.eu/events/replab2012](http://www.limosine-project.eu/events/replab2012)

In particular, given an input text in English, the model extracts: sentences and tokens, entity mentions (named entities, nominals, personal pronouns), syntactic information in form of constituent and dependency trees, opinionated expressions, relations between entity mentions, and co-reference chains. The user can run either parts of the pipeline or the complete system. An initial evaluation of this prototype has been carried out based on a community Question Answer dataset through the answer ranking task, i.e. finding the best answer to a question posed on a community-driven website. Results of this were published in SIGIR 2012.

The goal for the coming years is to extend the current semantic model to the other languages covered in the project (Spanish, Dutch, Italian) and identify cross-lingual mappings between semantic models. We also plan to extend our evaluation experiments to assess the impact of our deep integrated semantic representation on further information retrieval tasks.

**Semantic mining (WP4)**

The main objective of this work package is to mine meaningful patterns and properties of information objects relevant to the scenarios considered by LiMoSINe. In particular, new dynamic semantic search tasks are defined and the web of data is used to provide descriptions of patterns mined from text streams: around entities, events, online user behavior, and images. Information, text-based or semantic, pertaining to a single information object is aggregated to prepare for semantic aggregation search tasks.

In the first year, we mainly focused on semantic aggregation (in the form of entity modeling and entity linking) and on modeling automatic media annotation. The objective of semantic aggregation is to create a large knowledge base regarding information available in a set of
documents, such as webpages, blog posts, tweets, or community question answering websites. We aim at identifying entities in the documents and characterizing the connections among them. This knowledge will be used for the task of composite retrieval, where a user can query about an entity, or a small set of entities, and the system will compose a concise answer containing what is known about those entities. For modeling automatic media annotation, we have improved the performance of an existing state-of-the-art annotation model by exploiting tag co-occurrence and temporality. The first approach places preference to tags that frequently co-occur with suggestions made from an SVM. We were able to achieve statistically significant improvements to annotation accuracy on a collated Flickr collection.

The deliverables of WP4 in year 1 consisted of baseline versions of systems to address the tasks mentioned above. These include a semantic mining module, that offers facilities for automatically linking textual documents to structured knowledge bases in various languages; results were published at WSDM 2012. Additionally, a media annotation module for the automatic annotation of images using contextual information such as tag co-occurrence was produced as well as a semantic module for obtaining the historical interest and development around concepts; this module was released as a public API; see Figure

![OpenGeist home page.](http://www.opengeist.org)

**Access and recommendation (WP5)**

The main objective of this work package is to investigate novel techniques for effective access and recommendation of information by mining a set of disparate textual and semantic resources. The research work concentrates on data streams like Flickr (multimedia) and Twitter (Social streams) in the first year. A Flickr data set of 2M images and tags has been crawled and a demonstrator is developed. The work has been disseminated in a number of venues like ECIR 2012, SIGIR 2012, and WSDM 2012.

---

3 [http://www.opengeist.org](http://www.opengeist.org)
A module for cross-lingual information access was developed. The goal of this module is to offer novel techniques for cross-language information retrieval. It is based on semantic information, i.e., recognizing entities and perform a matching between queries and documents based on this form of semantic enrichment. We also propose a photo tag recommendation system, which automatically extracts semantics from visual and meta-data features, to compliment existing tags. Compared to standard content/tag-based models, these automatic tags provide a richer description of the image and especially improve performance in the case of the cold start problem. The results will be presented at ACM MMM 2013.

The paradigm of composite retrieval consists of extracting from a collection of available information items, a set of diverse item bundles that together form the best possible answer to a user's complex information need. We have provided a formal definition of the composite-ranking problem and developed efficient algorithms. The implementations of these algorithms constitute the initial nucleus of the composite ranking module that will be integrated in our entity-Driven Exploratory and sErendipitious Search SystEm (DEESSE) system, aimed to support a serendipitous exploration of complex data extracted from different social media. There has been growing interest in building aggregated search systems where information from a variety of sources (so-called verticals) is retrieved and aggregated into one single interface. We are developing a framework for evaluating as well as optimizing aggregated search systems in terms of relevance and diversity of information. Our framework is presented as a full research paper at SIGIR 2012. A diversity-based metric is presented at ACM CIKM 2012, where an aggregated search demonstrator is also being presented.

Our initial activities on context modeling and personalization focused on searchers and their behavior. We have studied searchers behaviors during an information retrieval process. We investigated the connection between relevance and brain activity using functional Magnetic Resonance Imaging (fMRI). In our initial study we measured the brain activity of eighteen participants while they performed four topical relevance assessment tasks on relevant and non-relevant images. The results revealed three brain regions in frontal, parietal and temporal cortex where brain activity differed between processing relevant and non-relevant documents—an important step in unraveling the nature of relevance and building better user modeling techniques.

### 3 Demonstrators

Many of the semantic aggregation technologies built as part of LiMoSINe will be integrated into three demonstrators. Important achievements are the full specification of the three demonstrators in terms of scenarios and test-beds, that allow us to showcase the new technologies developed as part to the projects, and the first prototyping of these demonstrators.

#### Photo Tag Recommendation

Despite decades of research the task of automatically “understanding” an image, identifying what types of objects it contains is still not possible for modern day computers. Most advancement in this area has been made in focused applications, e.g., in face recognition or in telling different landmarks apart. However, no system comes close to a general-purpose image labeller, covering anything from trees, to people, to buildings, to insects. Standard search systems therefore rely either on textual clues about an image content or on metadata,
such as the image date, its size or the person who uploaded it. Such information can sometimes be inferred, e.g., from anchor text of hyperlinks pointing to images, but more and more users also provide such information directly by tagging their photos. For this reason, websites hosting images such as Flickr encourage their users to provide tags for their photos.

Figure 3. Mining additional tasks through co-occurrence.

The purpose of tag recommenders is to assist users in choosing descriptive, relevant tags based on some available knowledge. Photo tag recommendation can be viewed as a retrieval problem with a composite query, which may include various properties of the photo, tags user already specified for it, etc. Special-purpose features mined through co-occurrences add a semantic layer to our initial knowledge about the photo. This is important especially in a case of a “cold start”, when user has not provided any initial tags. See Figure 3. Our system is built on a dataset of 1,857,460 images, which were crawled from Flickr. A user first uploads a photo, and its features are automatically extracted.

**Entity-Driven Exploratory and sErendipitous Search SystEm**

Unlike the familiar query-driven search paradigm in which relevant documents (or entities) are sought after, exploratory search has elements of uncertainty, when problem is poorly defined, or the information seeker is unfamiliar with the problem domain, or the task requires some exploration. A system supporting exploration and serendipity must provide semantically cohesive, yet interesting and diverse set of results, which would encourage the user to continue browsing and exploring. Our second demonstrator provides a high-level overview of data in the form of enriched entity networks, where entities are people, places, events, and so on, extracted from social-media content. Given a query entity, our system presents a subgraph around the query entity to the user for exploration.

Figure 4. Top 30 entities for Freddie Mercury in a Yahoo! Answers co-occurrence network.

We extract entity networks from two datasets (Wikipedia and Yahoo! Answers). A collaboratively edited online encyclopaedia, Wikipedia maintains a staff of editors, robots and an army of volunteers to maintain the quality of its articles. The curated nature of Wikipedia may make it a more trustworthy source of information. However, the freedom of
conversation on Yahoo! Answers may present its own advantages, containing within it the opinions, rumors, and social interest and approval.

**Online reputation management**

The LI&C demonstrator supports the task of online reputation management, both the monitoring and profiling purposes. The monitoring task consists in seeking and analyzing information related to the company, with the objective of detecting any topic that might damage its image. In contrast, profiling refers to a single or periodic revision of a company's reputation as it distils from news, opinions and comments expressed in social media or online press.

For the monitoring task, the system receives as input a stream of tweets containing the name of a company, and the goal is to (i) cluster the strongly related tweets in topics, and (ii) assign relative priorities to the clusters. For the profiling task, the system receives a set of tweets with the aim of (i) classifying them into related and unrelated with respect to a given company (filtering) and (ii) determining the polarity (positive, neutral or negative) of the related tweets. In order to determine if a text is related to a given entity we have implemented a vote mechanism that calculates a score depending on how many entity context words are found in the input text. The main idea of this method is to extract the WordNet concepts in a sentence that entail an emotional meaning, assign them an emotion within a set of categories from an affective lexicon, and use this information as the input of a machine learning algorithm.

The demonstrator permits to work with texts in English and Spanish. It is worth mentioning that the monitoring and profiling systems have participated in RepLab 2012, the first evaluation campaign for Online Reputation Management, achieving very competitive results. Moreover, to evaluate and analyze the adequacy of the proposed approaches, we have developed an interface that allows to easily compare the results obtained with those of the gold standard used in the RepLab campaign. This gold standard has been manually labelled by the reputational experts in LL&C.
4 Dissemination and exploitation

The first thing we did as part of our dissemination and exploitation activities to develop a communication plan. The main objectives of this plan are: to get qualified opinions, to engage our messages with qualified audiences (opinion leaders: journalists, authorities and bloggers), and to get media coverage. In addition, we prepared basic dissemination materials, including a poster, flyer and uniform presentation formats. These material are available on the LiMoSINe web site.

Actions to get qualified opinion

To reach the first objective, qualified opinion, we have created a web site and a Twitter profile (@Limosineproject) where we tweet the key messages of the project. It is currently pending creating a blog integrated on the website and a YouTube Channel for LiMoSINe Project. Both actions are scheduled for the second year of the project.

In addition, we have made a mapping with the specific people we are going to keep in touch with during the project. The third thing we did was to create a press kit with the main conclusions of the project including the analysis of Twitter and Spanish companies. The press helped us to disseminate the project and got media coverage. Additionally, we are making a video with interviews to the workpackages leaders. It will be finished before the end of the year.

Actions to get media coverage

Related to the second objective, we have participated in two types of events:

- **Scientific events:** LREC 2012 Workshop, CLEF 2012 Lab, the RAMSS workshop at ICWSM 2012.
- **Dissemination events:** workshop organized by Corporate Excellence (Centre for Reputation Leadership integrated by the most important companies in Spain). We talked about the LimoSINe Project to the leaders responsible for the online reputation in the IBEX 35 Top Companies.

Next year, we will be attending to the event called ‘Next International Conference on Corporate Reputation’, organized by Reputation Institute in Barcelona. In addition, we
established contact with some media in order to get coverage: El Confidencial, the main online media; El País, general information; and QUO Magazine, popular Science Magazine.

Looking forward, the project will be more attractive to the media when we have the first results of the research. Such details will allow us to make the LiMoSINe Project more interesting and increase the coverage in Spain as well as in other countries. Private sector and monitoring tool suppliers will be interested in the findings of the LiMoSINe Project. So, we could establish successful partnerships with them.