GALATEAS

Generalized **A**nalysis of **L**ogs for **A**utomatic **T**ranslation and **E**pisodic **A**nalysis of **S**earches

Annual Public Report 2011

http://www.galateas.eu

1. SCOPE

Part of the
European Commission 's
Information and Communication Technologies Policy Support Programme
(Co-funded by the European Commission for an overall budget of 3.7M Euros)
01/04/2010 to 31/03/2013

This summary is the annual public report of the Project GALATEAS (Generalized Analysis of Logs for Automatic Translation and Episodic Analysis of Searches). This summary covers the period 1 April 2010 – 15 November 2011.

2. GOAL AND CHALLENGES

Every day, millions of search queries are issued to content providers. These range from all-purpose web information sites (e.g. Google and Yahoo!) to digital library sites and merchant sites (e.g. Kelkoo and PriceGrabber). These queries are a precious resource in understanding user behaviour. From careful analysis of these queries, content providers can understand what information users are really looking for, the strategies they use to retrieve digital objects and the match between user needs and the content offered by the web site. Yet no provider has attempted to go beyond the standard log analysis service such as that provided Google analytics, to name but one.

*GALATEAS goal is to provide a **comprehensive analysis of user queries**, by identifying the information they contain from a **language interpretation perspective**, in a **multilingual context.***

To achieve that GALATEAS will address two important challenges:

- **Making sense of short queries in any language and**
- **Translating them.**

This will help content administrators to answer questions that are crucially important to them, such as:

- Which are the topics which are most commonly searched in my collection, according to a certain language?
- How do these topics relate with my catalogue?
- Which named entities (people, places) are more popular among my users?

3. GALATEAS SERVICES

To achieve the goal set above GALATEAS will provide two services, LangLog and QueryTrans.

3.1. LANGLOG

LangLog (Fig. 1) will analyze transaction logs of queries to search engines for a given content provider. In a way similar to that of standard log analysis systems which provide generalizations of paths of users inside a web site, LangLog will provide ***generalizations of the actions that information seekers perform based on language analysis***.

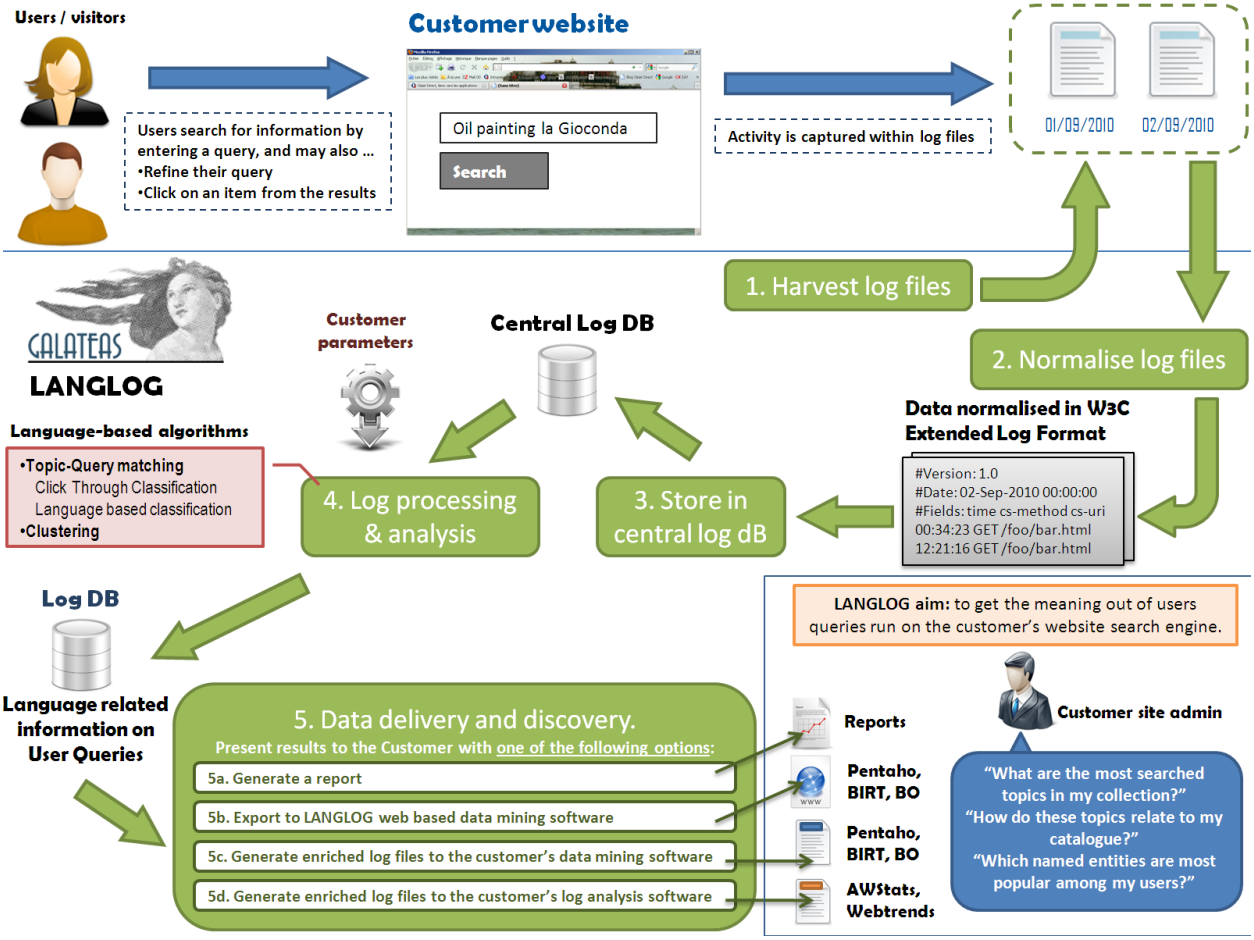

Fig. 1 High-level view of the LangLog service

LangLog will work with any log file generated by the customer's systems. The main processes involved in LangLog are:

- **Harvest** the customer's log files using the given access details, and running on a agreed period.
- **Normalise** the log files to the W3C extended log format, and **store the normalised data** in a central database (Central Log DB).
- **Process** the data from the Central Log DB using language based components (i.e. topic-query matching with Click-Through classification and Language-based classification, clustering) to extract information from the user queries before **storing** them in a Log DB
- **Present and deliver results to the customer**, either through a report, a web based data mining software, or by sending enriched log files to the customer's data mining or log analysis tool.

3.2. QUERYTRANS

QueryTrans (Fig. 2) will translate queries from an external search engine into several target languages. The external search engine will use these translations to return results in languages different from the one in which the query was formulated. QueryTrans will not be a cross language information retrieval system, performing indexing and search, but a ***query translation service***.

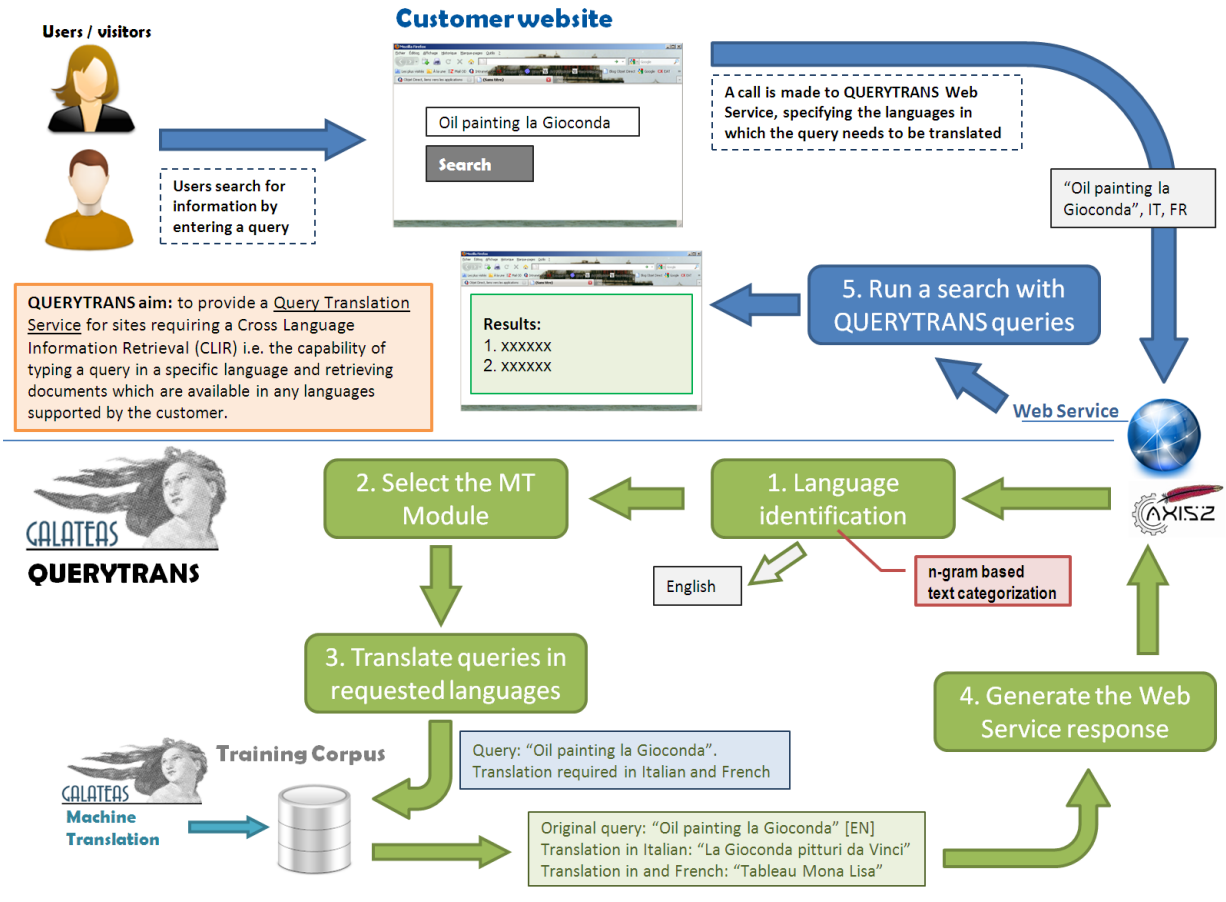

Fig. 2 High level view of the QueryTrans service

QueryTrans relies on a corpus (i.e. large and structured set of texts) produced by using LangLog data and generated from the customers log files. The main processes involved in QueryTrans are:

- Identify the language of the query.
- Select the most appropriate machine translation (MT) module for the domain.
- Translate the queries in the requested languages.
- Generate a web service response including the original and translated queries, which is sent back to the client's search engine where a search including the translated terms is launched.

4. PARTNERS

1. **Xerox Research Center Europe**

The European branch of Xerox global research**, Xerox Research Centre Europe (XRCE)** is a multidisciplinary organisation with experts in a broad spectrum of activities linked to information, data and documents.
Xerox is the administrative coordinator of the GALATEAS project and in charge of the technical development of linguistic tools and resources. Xerox also contributes to configuring the MOSES Machine Translation system and developing the QueryTrans service.

2. **CELI - Language and Information Technology**

CELI is a company founded in May 1999 by researchers and engineers, previously working in the research sector. The main goal of **CELI is to bring the results of the most advanced research activities in the Human Language Technology, Artificial Intelligence, and Web Interaction fields** into products and solutions for the enterprise.
CELI is the technical coordinator of the project. CELI also focuses on cross language and NLP technologies with a special focus on query logs alignment. Moreover it integrates skills and resources acquired in the context of CACAO project and facilitates the integration of language resources and algorithms from several partners.

3. **Objet Direct**

Objet Direct is a software, training and consulting company specialized in object / Internet technologies and distributed architectures
In GALATEAS Objet Direct is mainly involved in Log analysis, Data Mining and Integration activities. It also plays a crucial role in business and exploitation activities.

4. **Bridgeman Art Library - Digital Images of Art**

The Bridgeman Art Library(Bridge) is the leading international resource of digital images of art, artifacts, culture and history representing 2000 stakeholders in the museums, galleries and archives community worldwide.

Bridge is active in workpackages concerning Evaluation, Optimization and business sustainability. Moreover, being an international art portal and a multimedia digital library, Bridge provides the data for the tuning of the Click-through classification algorithm.

5. **Gonetwork srl**

GONETWORK s.r.l. is a company founded in 1999 with the goal of reaching a **good market position in the field of Web Marketing** and **Search Engines Optimization**.
In the context of GALATEAS, GONETWORK will contribute with its experience in transaction log analyis, web service integration and development of web Graphical User Interfaces.

6. **Universit degli studi di Trento - Italy**

The University of Trento unit consists of researchers belonging to the **Department of Information Engineering and Computer Science** (DISI) and the inter-departmental **Center for Mind/Brain Sciences.** (CIMeC). Most of the GALATEAS activities will be conducted by scientists from the CIMeC Language, Interaction and Computation laboratory (CLIC)**.** CLIC is one of the largest and most active centers for **human language technology** in Italy.
In GALATEAS UNITN is mainly active in the NLP and Algorithm WPs, with special attention to Named Entity Recognition, Clustering and Classification.

7. **University of Amsterdam**

The University of Amsterdam joins with researchers from the Informatics Institute (Faculty of Science). **The Intelligent Systems Laboratory Amsterdam (ISLA)** within the Informatics Institute focuses on processing information in pictorial, auditory and/or textual form.
The main tasks of UVA in GALATEAS concerns the adaptation of the MOSES MT system, and the tailoring of the various algorithm of log analysis.

8. **Berlin School of Library and Information Science Humboldt-Universität zu Berlin**

The School of Library and Information Science of the Humboldt-Universität zu Berlin (IBI) is the oldest school of library science in Germany, the only library school at a research university, and the only German institution with the right to give a doctorate in library and information science.
Tha main tasks of UBER in GALATEAS are related to providing data and evaluating different GALATEAS components.

5. SUMMARY OF THE OBJECTIVES AND WORK CARRIED DURING THE REPORTING PERIOD

The project started officially on the 1st April 2010 and is going to run for 36 months. From April 2010 to November 2011 the project has developed one of the main project services, LangLog,

and configured a widely used statistical machine translation system, MOSES, to deal with queries and to easily adapt to different application domains. The activities undertaken in the project up to now can be summarized as follows.

1.1.1 LANGLOG SERVICE – THE FIRST SERVICE DEDICATED TO LANGUAGE-BASED QUERY ANALYSIS

Developing LangLog, the service and infrastructure enabling language analysis of query transaction logs and the generation of subsequent customized reporting, has been the main technical objective of this period. This is a highly complex and challenging task, requiring the following components and services.

- Language Resources and Services: The components, tools, and resulting services that provide all language analysis functionalities. These help answer questions such as what are the languages in which users query the system, using for example a language identification service, or which named entities are mostly searched for, using a named entity recognition service. Other services developed include query tokenization, lemmatisation, and disambiguation, a dictionary lookup, and semantic similarity. The linguistic information and language identification are available for French, Italian, English, German, Dutch, and Polish. Named entity recognition is available for English, French, and German.
- Clustering and Classification Services: Being able to identify similar objects such as queries can help in identifying the type of objects that users are looking for and adapt/modify ones collection or data structures. For example, if based on the query logs the words “banana”, “pineapple”, and “cherry” form a cluster this indicates that they could be grouped together. For the cases that the digital content provider has a hierarchy, GALATEAS has also a classification service that assigns the queries to categories from the provider’s hierarchy, revealing patterns of user search relevant to the provider’s internal data storage structures.
- Log Management Components: A dedicated framework, the Pentaho Extract Transform Load (ETL) one, has been used to set up the log management workflow. This allows dealing with different harvesting requirements through customized scheduling and various log formats by applying appropriate normalisaiton functionality. Customisations of the ETL in the context of GALATEAS allow also the enrichment of data using the services mentioned above.
- Log Disclosure Components: Normalising and enriching data with all the extra language related information is not enough. Providing this data to the customer in a human-friendly way is also required. The QlikView tool was selected for customized reporting, providing to the GALATEAS customers dynamic visualization capabilities.

Besides QlikView, the Pentaho platform is also being configured to enable the periodic generation of static (PDF) reports. A plug-in for AWStats (a widely used log analysis system) was also developed. The developed plug-in adds a menu entry in the AWStats standard menu. Through this additional menu, the user can view a report about query enrichment along with some charts.

The LangLog services website shown in Fig. 3 and found at http://langlog.galateas.eu/ includes more information on LangLog, illustrates some analysis examples, and gives the opportunity to try some interactive demos.

Fig. 3 Screenshot of the LangLog services website

1.1.2 QUERY TRANSLATION SERVICE (QueryTrans service)

In the context of GALATEAS a dedicated service to translate queries in the different languages of the project is being developed. Such a service would enable the analysis of the needs of digital content providers considering information in different languages and will also allow users to perform multilingual searches. The MOSES statistical machine translation system has been selected as the backbone of that service. During the reporting period it has been configured to deal with short queries and to be easily adaptable to different application domains.

A dedicated software called TLike links LangLog to the translation part of GALATEAS. It transforms the data available from LangLog to a form that allow translations to be adapted to the requirements of new domains (and potentially clients) in terms of the language being used.

The QueryTrans service that will integrate these developments is being implemented and a first version will be available at the beginning of 2012.

6. SCIENTIFIC INNOVATION

Mainstream services in the field of web logs are click rate, visited pages and user paths inside the document tree. The GALATEAS services will analyse the information contained in queries from the point of view of language interpretation, not data. Towards that objective a number of language services are already available. However, making sense of short queries is particularly challenging. The consortium is developing new approaches to deal with this challenge.

GALATEAS has implemented a web service, LangLog, that provides a view of queries into conceptual units that allow administrators and managers to answer questions such as: "What are the topics most commonly searched in my collection, in a certain language?"; "How do these topics relate to my catalogue?"; "Which named entities (people, places) are most popular among my users?". This is done through customised log analysis, query clustering and topic computation methods that consider language information (e.g. named entities and distance between semantic vectors). Some example scenarios and demos can be found on http://langlog.galateas.eu/.

In machine translation, GALATEAS investigates the use of statistical machine translation systems to provide meaningful results for short, syntactically and contextually poor texts such as queries to search engines. The backbone of the technology used is the Statistical Machine Translation system, MOSES. Different tuning approaches are proposed in the project to deal with query and domain adaptation. Based on this technology GALATEAS will provide a

dedicated service to query translation that will be adaptable to different domains (e.g. medicine) according to the needs of the service users.

7. EXPECTED FINAL RESULTS

Galateas will offer two types of services to content providers:

- **LangLog** will extract meaning out of lists of queries for library, content aggregator and site managers.
- **QueryTrans** will be the first web translation service specifically tailored to query translation.

Languages addressed by LangLog and QueryTrans will be Italian, French, English, German, Dutch, Modern Arabic and Polish.

8. TARGETED USERS

Indirect "users" of the service are information seekers, who can benefit from improved, possibly cross lingual, search services. The GALATEAS services are not however provided directly to end users, but to administrators and managers of content aggregators and search engine installations. GALATEAS will therefore target a high end B2B market where customers will be mostly represented by organizations running middle and large sized aggregated content.

The main market segments targeted by GALATEAS are: Digital Libraries, Content Aggregators, and Merchant Sites.

9. POTENTIAL IMPACT

GALATEAS will provide the first query oriented machine translation system. GALATEAS will therefore aid in overcoming language barriers across Europe by allowing any user speaking one of the GALATEAS languages to type a query in his/her mother tongue and retrieve documents/metadata in several languages.

GALATEAS other crucial contribution relates to multilingual web content management. By providing support to query log analysis for at least seven languages, GALATEAS will enable content providers to understand what their users are looking for (contrasted with what they found, which can be computed simply by counting user's click) and in which language: this means, for content providers, the possibility of a much more targeted acquisition of new contents; for users, this will imply access to more pertinent contents.

Software developed in the context of GALATEAS that is subject to open source licenses will be made available under open source licenses. For example, GALATEAS makes use of open source

software programs, such as MOSES. GALATEAS will therefore enable open access to source code and systems and will support community-based collaboration and evaluation by running a query translation experiment in the context of CLEF.

10. MORE INFORMATION

More information can be found on the project's website in multiple languages (URL: http://www.galateas.eu/). The website includes a two-pages summary factsheet (URL: http://www.galateas.eu/files/galateas_4.pdf), and a powerpoint presentation (URL: http://www.galateas.eu/files/CIP_250430__1_0_CONT_CORE_2.pdf) that give you a quick overview of the project. Do not hesitate to refer to the logos of the project shown below.

11. CONTACT DETAILS

Coordinator: Xerox Research Centre Europe (France)
Xerox Research Centre Europe
6 chemin de Maupertuis - 38240 Meylan - France

URL:http://www.galateas.eu/