Publishable summary

Grant Agreement number: 250467

Project acronym: ATLAS

Project title: Applied Technology for Language-Aided CMS

Project type: ☐ Pilot A ■ Pilot B ☐ TN ☐ BPN

Period covered: 30.11.2010 - 30.11.2011

Project coordinator name, title and organisation:

Anelia Belogay, CEO, Diman Karagiozov, CTO,

Tetracom Interactive Solutions

Tel: +35924950444

Fax: +35924950443

E-mail:anelia@tetracom.com, diman@tetracom.com

Project website address: www.atlasproject.eu

Introduction

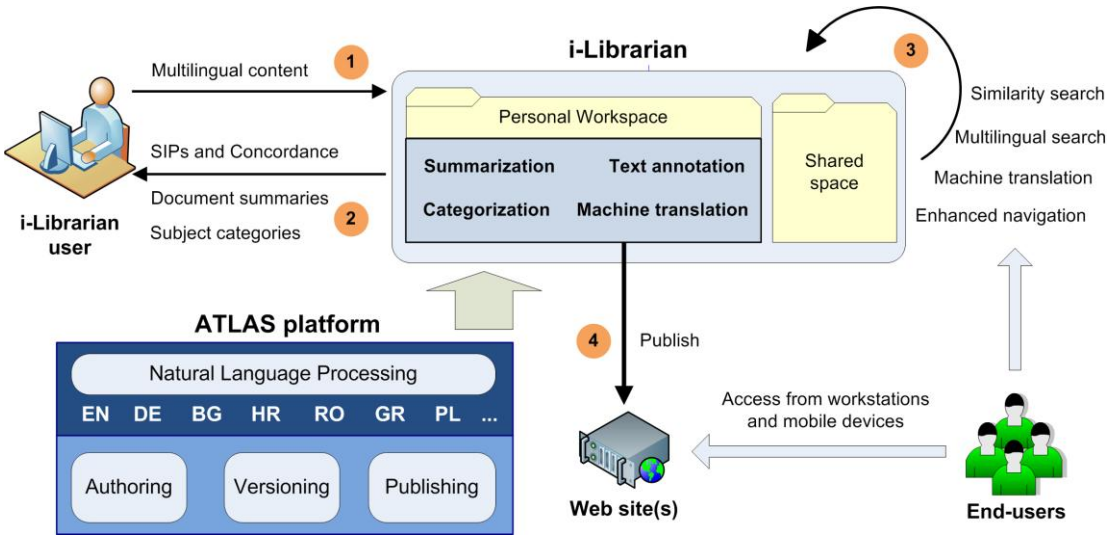
The advent of the Web revolutionized the way in which content is manipulated and delivered. As a result, digital content in various languages has become widely available on the Internet and its sheer volume and language diversity have presented an opportunity for embracing new methods and tools for content creation and distribution. Although significant improvements have been made lately in the field of web content management, there is still a growing demand for online content services that incorporate language-based technology. Mechanisms such as automatic annotation of important words, phrases and names, text summarization and categorization, and computer-aided translation could facilitate the process of manipulating heterogeneous multilingual content as well as enhance end-user experience by allowing for better content navigation. This project unifies such mechanisms in a common software platform called ATLAS and builds three separate solutions around this platform.

The project solutions

i-Librarian – the intelligent content assistant service

The first solution, i Librarian, is a web-based content assistant service, which allows users not only to store, organize and publish their personal works but also to locate similar documents in different languages and to obtain easily the most essential texts from large collections of unfamiliar documents or search engine results.

i Librarian is a web-based content assistant service, which encourages visitors to register and get a personal workspace where they can store, share and publish various types of documents and have them automatically categorized into appropriate subject categories, summarized and annotated with important words, phrases and names. Advanced language-based technology is implemented to help users easily navigate between and access both their personal works and unfamiliar documents. After processing a large collection of unfamiliar texts i Librarian displays short summaries and extracted concepts that enable users to easily decide which documents are worth reading and which could be discarded. Furthermore, i Librarian interlinks all user documents based on the extracted phrases, words and names, and thus improves significantly content navigation. Finally, the service helps users with no previous experience to publish their own content using the power of a modern content management system but without struggling with the inherent complexity of such systems. The features of i Librarian will be initially available in seven languages – English, German, Bulgarian, Croatian, Greek, Polish and Romanian. However, as more languages could be easily integrated in the service, the consortium will explore several options to secure the necessary funding after the end of the project for supporting all other major European languages.



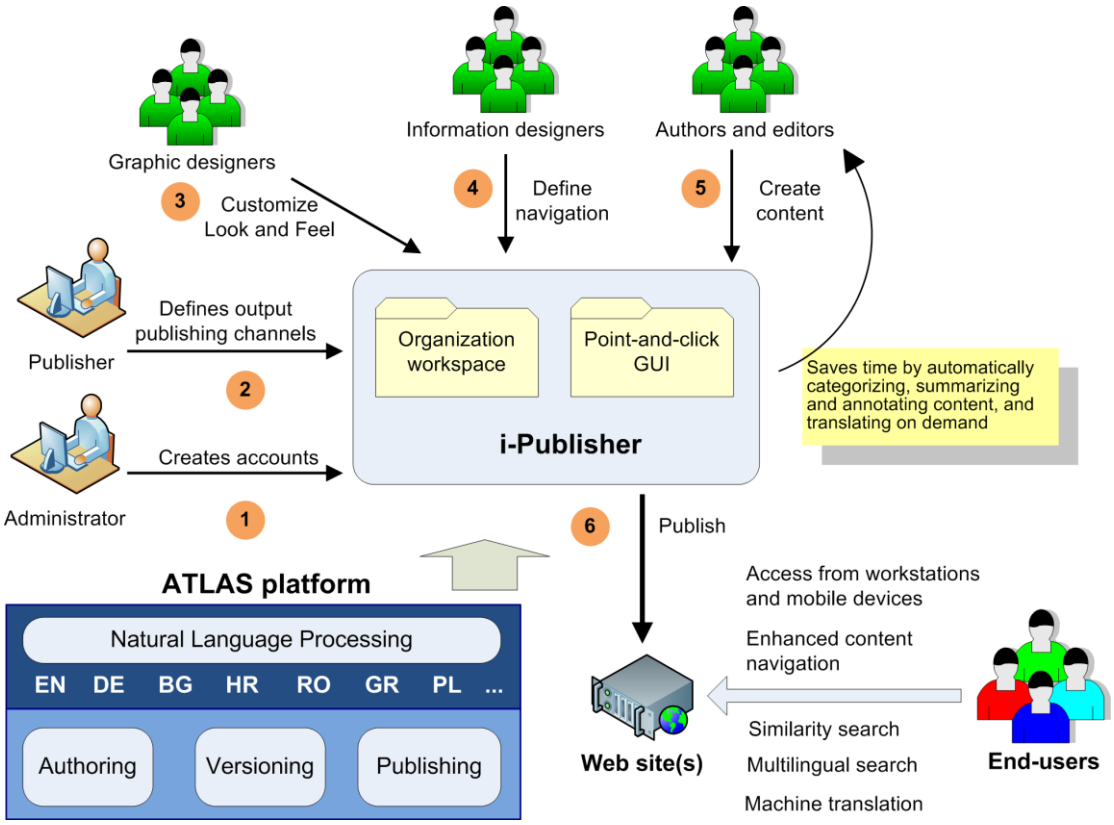
Some of the main characteristics of i Librarian are summarized below:

- i Librarian offers multilingual full-text search inside personal or shared documents.
- i Librarian provides multilingual similarity search, which enables users to easily locate both personal and shared similar documents in different languages.
- i Librarian implements powerful instruments for computer-aided translation, automatic content categorization, summarization, and annotation of important words, phrases and names.
- Users can rate the quality of automatic translations and improve them, which would help the consortium to build better translation models for future use.
- Users can publish heterogeneous multilingual content on a personal web site hosted by i Librarian or on existing web sites and portals.
- Users can freely annotate documents in their personal workspace and search through the annotations (possibly in different languages) shared by other users in order to find documents of interest.
- i Librarian is accessible from a browser or mobile devices such as iPhone.
- i Librarian includes a mechanism for reporting and removing of materials that violate copyright laws.

i-Publisher – the online web content management solution

i Publisher is a novel software-as-a-service solution for web content management, which allows both small and large organizations to deploy and manage multilingual web sites without spending time and efforts for installing and maintaining a content management system. This service assists organizations in retrieving, unifying, and packaging heterogeneous pieces of content, and dynamically rendering them on multiple web sites. i Publisher fosters collaboration in content creation by enabling authors, editors, and other

contributors to work together. It also facilitates the process by automatically categorizing, summarizing, and tagging the newly created content. Furthermore, web sites may be built with i Publisher with a point-and-click graphical user interface by people with different expertise but no programming experience – publishers, information designers and graphic designers. The service leverages the full benefits of the ATLAS platform and becomes an ideal choice for promoting any type of organization on the Web. i Publisher will be available free of cost for non-commercial use in order to promote web standards and encourage language diversity in content creation. Different subscription plans will be available to those who desire more storage space and customer support, or who would like to use i Publisher for commercial purposes.



i-Publisher characteristics :

- i Publisher is well-suited to both small and large organizations as it is designed with scalability in mind, i.e. if an organization needs to handle more content and users, additional servers will help address these needs and provide the desired results in terms of performance.
- i Publisher improves content navigation by dynamically interlinking content items based on extracted important words, phrases and names.
- i Publisher utilizes a flexible user access rights system comparable to that of a modern server operating system – security policies may be set for groups and specific users as well as for specific content items or even content item properties.

- i Publisher implements an industrial strength versioning system, which supports the versioning of structured content rather than the simple text-based versioning found in most existing solutions.
- i Publisher allows content to be mass exported to or imported from file systems, databases or file servers.
- Web sites created with i Publisher offer to end-users multilingual full-text and similarity search as well as clustered, summarized and annotated content.

Summary description of project objectives

The consortium will adjust and integrate several existing software components, assembling a platform for multilingual web content management called ATLAS, and a visualization layer called i-Publisher, which adds to the platform a powerful web-based point-and-click tool for building, reusing and managing multilingual content-driven web sites. An instance of i-Publisher will be made publicly available as an online service. i-Publisher will also be used to build two thematic content-driven web sites – i-Librarian and EUDocLib.

The ATLAS project aims to meet the following objectives:

- Software platform and services, demonstrating the latest achievements in the field of multilingual web content management and addressing the needs of individuals and organizations for easier web site building and content publishing.
- Liaison with the Europeana and EuroMatrix Plus initiatives in order to foster language diversity in content creation and distribution
- Interoperability by conforming to a number of widely recognized web, natural language processing, and content management standards
- Sustainable management format to ensure the progress of the project
- Mechanisms and procedures that enable and simplify the addition of new languages to the ATLAS platform, thus targeting all major European languages after the successful completion of the project.

Description of work performed since the beginning of the project and main results so far

With regard to the management objectives set for the first period the following tasks have been completed:

- A management and coordination framework was established to ensure the smooth progress of the project.

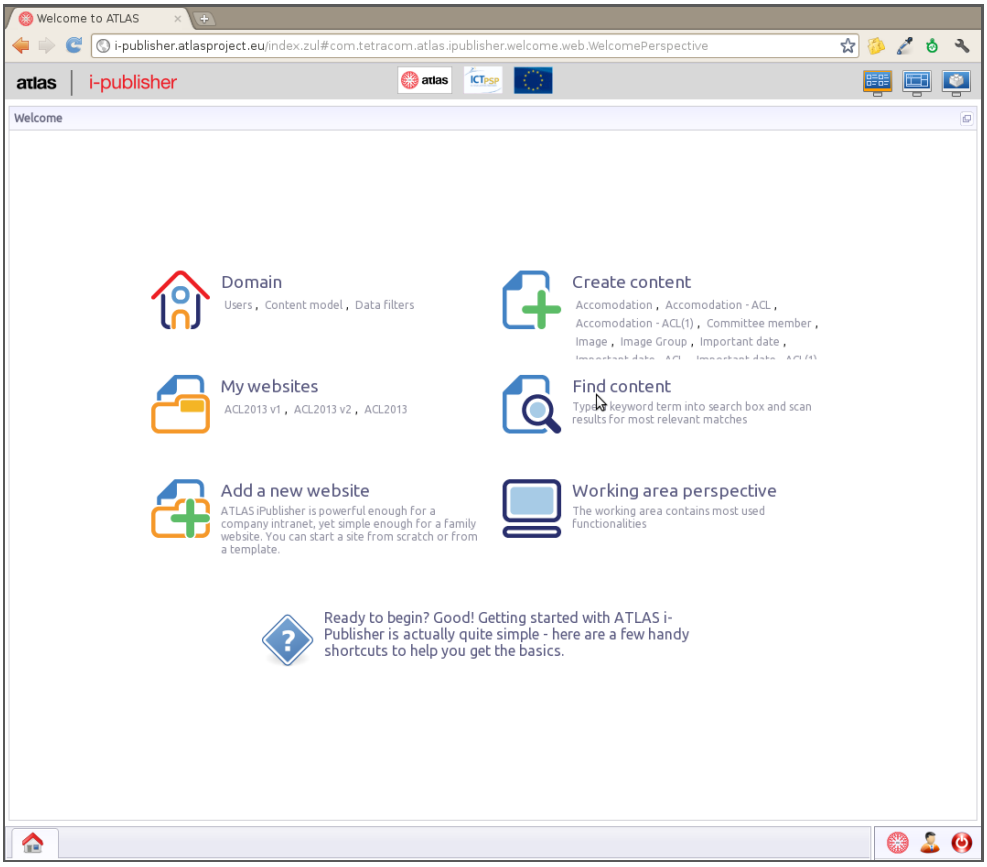
- The consortium agreed on a process through which to monitor the allocation and distribution of project resources, as well as to control the quality and timely delivery of project deliverables.
- Seven project meetings were organized (the kick-off and two WP meetings.) A common understanding of the project goals was gradually achieved on these meetings. Furthermore, the consortium was able to smoothly define the next steps needed in order to achieve the objectives for the next period.
- Channels ensuring the good management and technical communication were established.
- The first periodic report covering month one through month six of the project was prepared and submitted to the EC.

The work done in terms of the technical objectives set for the period includes the following:

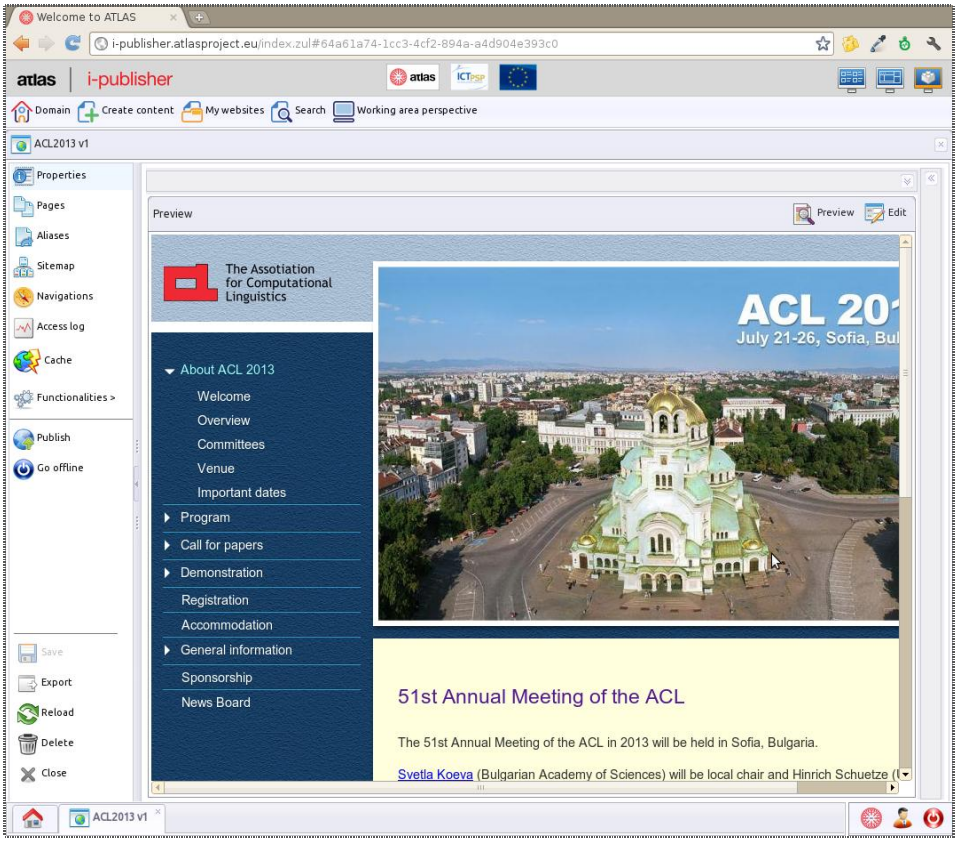
- **i-Publisher** was built and integrated into ATLAS CMS
- Two on-line services **i-Librarian** and **EUDocLib** were built with i-Publisher
- **i-Publisher Simple mode** was built to help inexperienced users creating their website by using ready-to-use websites and templates

http://i-publisher.atlasproject.eu
http://i-librarian.eu
http://eudoclib.atlasproject.eu

i-Publisher

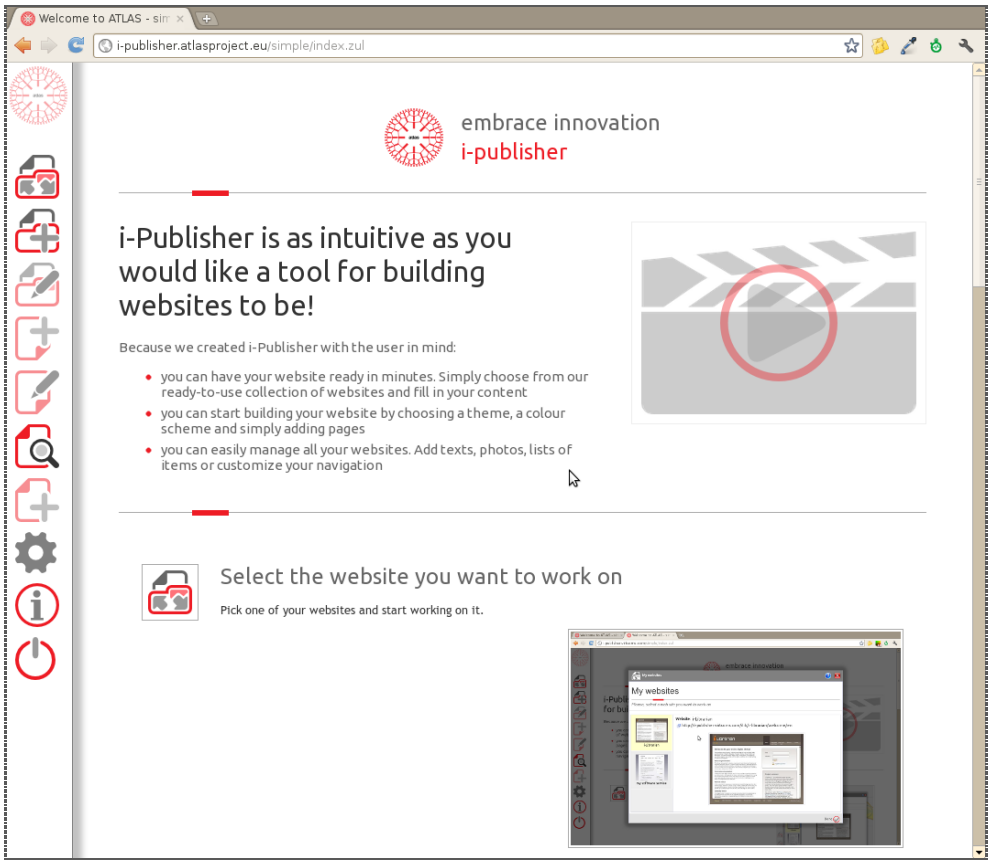


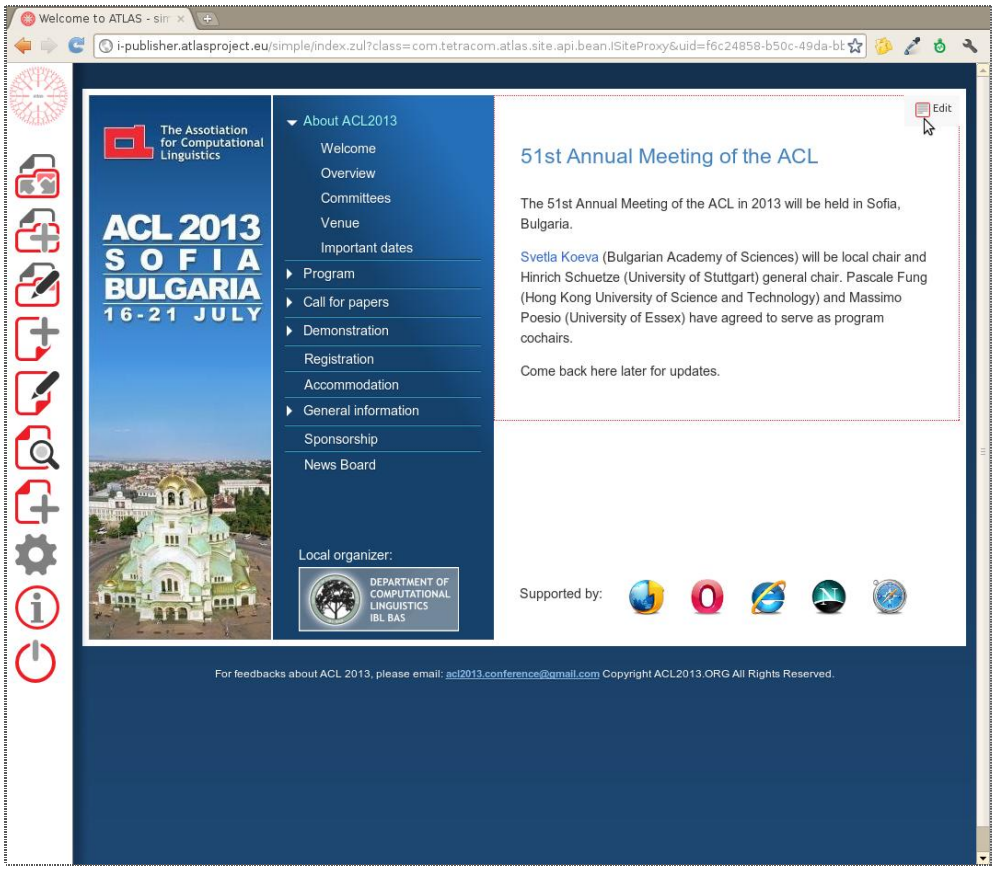
Home page



Edit a website

i-Publisher Simple Mode





Edit a website

i-Librarian

Book details page

Home page

EUDocLib

Home page

Search result

- Improvements of usability of i-Publisher – export/import and reuse of a web site; various functionalities were extended in order to increase the productivity of the users.
- The first drafts of the user guides of i-Publisher Advanced Mode and i-Librarian are accessable. The implementation of the technical documentation of the three services has started.
- The categorization engine was extended to support several algorithms, namely Relative Entropy, Naive Baysean, Class-Featured Centroid, etc. In addition, the engine now supports combining of different classifiers which results in better predictive performance.
- The LPCs implementation of the language processing chains (LPCs) for all project languages was carried out, targeted at providing tools for sentence-splitting, tokenization, lemmatization, POS tagging, NP extraction and NE recognition. The Bulgarian, Polish and Greek LPCs are integrated in Atlas and German and Romanian are in a test phase.
- Two different summarization approaches that will provide the user with a summary depending on text length were defined. Two long-text summarization methods were developed: a "shallow" one and another that exploits the full chain of summarization tools – Language Processing Chain, RARE (Robust Anaphora Resolution Engine), Discourse Parser, clause splitter and summarizer.
- The 1st round of the ATLAS system User Acceptance Evaluation has been conducted and the results were presented to the Consortium and the Commission. As a result, additional simplified mode was added to the functionality of i-Publisher, the quality of the work of the English chain was improved.
- The MLeCeL LL was established in Sofia and a website is accessible at http://livinglab.itd-bg.eu/ . The two host organization teams set their laboratories and started conducting experiments with the services.
- Technical Indicators for the CMS, summarization, machine translation, Cross-lingual Content Retrieval have been drafted.

Work carried out on dissemination for this period includes:

- The new project website was created with i-Publisher.
- Several major dissemination events were held:
  - ATLAS Workshop "Integration of multilingual resources and tools in Web applications workshop" @ Conference of the German Society for Computational Linguistics and Language Technology (GSCL 2011)
  - ATLAS – Multilingual Language Processing Platform paper presentation at SEPLN 2011
  - The ATLAS project and i-Publisher, i-Librarian were presented and demonstrated in a stand at WEBIT 2011 (webit.bg) to a wider audience (5000 visitors, 40 countries)
- A new vision of the project and its services is applied to the dissemination materials and services websites.
- The project has been presented to linguists, computational linguists, NLP engineers, IT scientists, key players in Language Resources and Technologies and representatives of social sciences and humanities, business people, students and researchers in conferences, meetings described in details on the project website.

Expected final results

The primary goal of the ATLAS project is to facilitate organizations and individuals who manage and publish multilingual content. Thus, the project solutions will not merely meet the needs of modern multilingual content management, but also create value for all users.

Main expected final results:

- The software solutions built during the project reveal the true value, capabilities and power of several existing tools for web content management, multilingual versioning, and natural language processing by combining them in an innovative manner and offering the end results to the general public at no cost.
- With i-Librarian and i-Publisher users can easily create, manage and publish multilingual content without installing and maintaining a standalone system. Nevertheless, they retain full control over their content regardless of whether it is in their private workspace, shared or published. EUDocLib provides easy and intuitive access to a vast collection of EU law documents.
- The ATLAS platform is designed with extensibility in mind, which allows for easy addition of tools for currently unsupported languages as well as new tools for already supported languages.
- Furthermore, ATLAS significantly reduces the time and efforts for content authoring and editing because it automatically categorizes, summarizes, annotates and translates documents regardless of their language and format. The software platform enables i-Librarian users to find the most essential texts from large document collections by displaying text summaries and extracted important phrases, words and names.
- Finally, ATLAS improves content navigation by interlinking content items based on text annotations and by automatically placing the content items in appropriate subject categories.

Potential impact

The project brings together advanced technologies for multilingual web content management and text mining (such as automated annotation, mark-up and translation) in a united platform. The intended software-as-a-service architecture of the envisaged solutions, which demonstrate the capabilities of the ATLAS platform, and the open-source license, will facilitate the spread of the project output.

Main expected impacts:

- Technological

- Integration of text mining tools into content management systems
- Integration of text mining services
- Stable and more efficient Machine Translation modules for the project languages. The language pairs considered in ATLAS are covered by Google Translation but with very low quality. On the other hand these language pairs have strong relevance for the Central- and East-European commercial space.
- Contribution to the development of text processing chains for languages, which lack resources at present
- Adherence to and promotion of existing and future web standards
- Practical and economically viable solutions for nearly-automatic provision of multilingual online content and services for some EU languages

- Social
  - Facilitate exchange of information and knowledge
  - Simplify authoring, management and exploitation of heterogeneous multilingual content
  - Address the needs of a large number of people belonging to different target user groups – individuals and organizations
  - Cross the language barrier
  - Facilitate culture exchange
  - Liaise with Europeana and EuroMatrix Plus – The liaison with EuroMatrix Plus will be established at the beginning of the project. Europeana will be approached by the end of the first year, when the consortium will be able to demonstrate the potential value of ATLAS to the European digital library.

Use

The ATLAS platform as a whole and also some of its standalone components are beneficial to different groups of users. Thus the consortium has distributed the potential users of each major software component into several target groups while paying special attention to the needs and requirements of each group. The table below summarizes this distribution:

Target groups

Component	Target group
ATLAS (includes KMS Content Management System, Text Mining engine, Search engine, Machine Translation engine) + **i-Publisher** (ATLAS web-based graphical user interface for building interactive, content-driven web sites)	Web design companies – faster prototyping, web design and site building
	Hosting companies – as part of hosting packages
	Education, Media, Publishing, Non-profit, Government
Text Mining engine	Online bookstores
	Digital libraries/repositories
	News agencies/websites
i-Publisher (as online public service)	Small enterprises
	Non-profit organizations
i-Librarian (thematic content-driven web site built with i-Publisher)	Students, Researchers
	Readers
EUDocLib (thematic content-driven web site built with i-Publisher)	The general public

Table 1: Target groups

More information including project details, news, and contact information can be found at:

www.atlasproject.eu