# PROGRESS REPORT

**Grant Agreement number:** 250467

**Project acronym: ATLAS**

**Project title:** **Applied Technology for Language-Aided CMS**

**Project type:** ☐ **Pilot A** **V Pilot B** ☐ **TN** ☐ **BPN**

---

**Periodic report:** **1st ☐ 2nd ☐ 3rd V 4th ☐**

**Period covered:** **from 01.03.2012 to 28.02.2013**

---

**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**

# DECLARATION BY THE PROJECT COORDINATOR

I, as coordinator of this project and in line with my obligations as stated in Article II.2 of the Grant Agreement declare that:

- The attached periodic report represents an accurate description of the work carried out in this project for this reporting period;

- The project (tick as appropriate):

  V has fully achieved its objectives for the period;
  has achieved most of its objectives for the period with relatively minor deviations;
  has failed to achieve critical objectives and/or is deviating significantly from the schedule.

- The public Website is up to date;

- [this point only applies to projects with actual cost reimbursement] To my best knowledge, the information contained in the financial statement(s) submitted as part of this report is in line with the actual work carried out and consistent with the reported resources and if applicable with the certificates on financial statements.

Name and position of Coordinator: Anelia Belogay, CEO

Date: 31.03.2013

Signature:

# PUBLISHABLE SUMMARY

### 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 summarisation and categorisation, 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.

## 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.

## 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, shared or published workspace. 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 automatic 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
  - Integration of a cross-lingual search engine based on semantic web principles
  - 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 –Preparation for further integration of Mt-Modules with EUROMatrix Plus.

## 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 CMS Content Management System, Text Mining engine, Search engine, Summarisation Engine, Machine Translation engine, Cross-lingaul search 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

### Project web site

More information including project details, news, and contact information can be found at http://www.atlasproject.eu

# PROJECT PROGRESS

## Project objectives for the period

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

- The management and coordination framework ensured the smooth progress of the project.
- The control of the quality and timely delivery of project deliverables as well as monitoring of the allocation and distribution of the project resources resulted in implementation of all tasks as planned.
- Two project meetings were organized during the reported period. The first one was held in Luxembourg. The Consortium checked all documents related to the upcoming review meeting and discussed the project progress and defined the next steps needed to achieve the objectives for the next periods. The second meeting was held in Sofia and was focused on the preparation of the final test round and the dissemination events and activities organized at the end of the project by all partners.
- The second year report covering month twelve to month twenty four of the project was prepared and submitted to the EC.
- The second review meeting was held in Luxembourg.
- The periodic report covering the development of the project for months 24 to 30 was delivered to the Commission.
- The public report was delivered to the Commission.

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

- WP 2
  - ATLAS improvements – major performance and stability improvements; improved workflow of the process of invitation and user registration.
  - i-librarian extension made possible the internal evaluation of ATLAS on corpus containing 186 manually annotated documents. The evaluation results can be accessed at: http://www.i-librarian.eu/ilib/i-librarian/eval_list using the following user account: username: eval@i-librarian.eu, password: eval1234
  - Regression test module – a strategy for regression tests was built as a result of the reviewer recommendations. It was followed by the implementation of a regression test module.
  - An integration module for one of the most popular open-source content management system Alfresco was done. The module provides the linguistic annotations in XML and/or JSON format; the annotations are parsed by the implemented Alfresco extensions.
  - The technical documentation was updated with the newly integrated modules (categorization and summarization).

- Integration of machine translation and cross-lingual search engines

- WP 3
- The final fine tuning of the categorization tool was done.
- A standalone evaluation tool for finding the optimal model settings was implemented. Using the tool we achieved results above the state-of-the-art reported figures on the Reuters-21578 corpus.
- Using the categorization tool we found the optimal parameters of the classification models in i-Librarian in all project languages.
- A new approach to categorise content by using metrics to measure the similarity between content items and category vectors was implemented and applied in the demo services.
- The Deliverable D 3.1 was sent to the Commission.

- WP 4
- Strategy for regression testing has been prepared and adopted. Respective document has been sent to the EC on 24 May 2012 together with ATLAS assessment after performing the internal evaluation.
- A multilingual resource GEO-Names (ATLAS-GEO) which provides content similar to JRC-Named content focusing on geographical locations has been created and all LPCs extended with it in order to increase the quality and recall of the extracted locations.
- Several NER engines have been combined in order to improve the quality of the extracted person names and organizations.
- Freebase person name NER has been implemented using lexicons built form the Freebase lists of people.
- BOM UTF-8 text extraction bug has been found and fixed.
- Bug fixes and minor improvements in all LPCs – EN POS tagger and lemmatizer, EL and PL NP extractor, BG NER, DE stability (see more details in task descriptions below).
- Extension of the LPCs to provide richer linguistics annotations for the summarization tool in WP5 – e.g. gender information for nouns, improvements in NP annotation etc.

- WP 5
- Termination of the activities related to the acquisition of the corpora to be used in the summarisation system training and evaluation;
- Of-line testing of all modules in the summarisation chain;
- Improvements addressing the functioning of the language specific modules;
- Integration of the modules in the ATLAS framework;
- Testing and evaluation of the integrated system for each language;
- Final adjustments and improvements;
- Writing the technical and scientific documentation.

- WP 6

  **Machine Translation**

  - Set-up of a methodology for acquisition of domain specific parallel corpora for all language pairs
  - Collection / automatic generation of parallel corpora for all language pairs
  - Set-up of a methodology for domain adaptation
  - Training domain adaptation models for all language pairs and 13 domains
  - Data preparation for example –based machine translation engine
  - Integration of SMT and EBMT in the ATLAS System
  - Evaluation
  - Documentation

  **Crosslingual retrieval**

  - Finalization of the communication interface between the cross-lingual retrieval engine and the ATLAS System
  - Integration of the Crosslingual research engine into the ATLAS System
  - Monolingual and Multilingual tests with the CLIR
  - Evaluation
  - Documentation

.

- WP 7
  - Finalisation of the 2nd and 3rd rounds of the ATLAS system User Acceptance Evaluation; statistical analysis and drawing of conclusions.
  - Finalisation and submission of the D7.2 “Analysis of User Evaluation”.
  - Integration and regression testing of the whole ATLAS platform, as well as technical evaluation and comparative assessment of the main platform components (LPCs, CMS, categorisation, MT, CL-IR, and summarisation).
  - Finalisation and submission of the D7.3 “Final Report on Test Results”.
  - Demo services and prototypes. Several demo services were implemented to prove the qualities of ATLAS services and to be used as main dissemination instruments.

- WP 8 – work carried out on dissemination for this period includes:
  - ATLAS Demonstration in five large exhibitions and conferences
    - ATLAS stand at the biggest world digital fair CeBIT 2012, http://www.cebit.de/product/atlas-build-your-website-at-no-cost/291713/C913812

    - ATLAS services were demonstrated at the Demo section at the most significant scientific event for the year in Europe – EACL 2012, http://eacl2012.org/system-demonstration/index.html
    - ATLAS was presented with a paper at the biggest linguistic conference LREC 2012, http://www.lrec-conf.org/lrec2012/?List-of-accepted-papers
    - The project was presented with a stand at META-FORUM 2012
    - The project presentation was held at CESAR META-NET Roadshow, Sofia, http://dcl.bas.bg/en/cesar_roadshow_program_en.html

- ATLAS and its services were presented in an interview on the Bulgarian National Radio: (http://www.binar.bg/%D0%9F%D1%80%D0%B5%D0%B4%D0%B0%D0%B2%D0%B0%D0%BD%D0%B8%D1%8F/%D0%A2%D0%95%D0%A5%D0%9D%D0%9E-%D0%A0%D0%90%D0%94%D0%90%D0%A0.html)
- Twelve ATLAS demonstration websites produced and updated for exploitation purposes

  Nine meetings for possible exploitation of ATLAS of Tetracom with WHO, UNESCO, Bulgarian National TV, M3 Communications Group, NATO, ISN at ETH Zürich, Foundation "Educational programmes".
- Twenty two publications (articles, papers and books) at International conferences and workshops
- One conference (by UHH) and three special workshops (by ITDF) were organised
- Poster presentations in major conferences – EACL 2012, EAMT 2012
- Four user evaluation workshops organised by ITDF
- Five seminars organized by academic partners within universities
- Continuing synergies with META-NET FP7 Network of Excellence and CESAR – Central and South-East European Resources
- Performed dissemination activities during the reported period (01.03.2012-28.02.2013) are summarised in Annex 1.