



PLUTO

Patent Language
Translations Online

Annual Public Report November 2010



The PLUTO project has received funding from the European Union's ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme, CIP ICT-PSP under grant agreement n° 250430

Document Status

Project	PLUTO: CIP ICT-PSP Grant # 250430
Document version	2.0 FINAL
Primary Author	Dr. John Tinsley, DCU
Review	Dr. Páraic Sheridan, Prof. Andy Way
Status	Public

TABLE OF CONTENTS

1. EXECUTIVE SUMMARY	4
2. PROJECT OVERVIEW	5
<i>Consortium.....</i>	<i>6</i>
3. PROGRESS TO DATE	7
<i>Achievements</i>	<i>7</i>
<i>EPO Collaboration</i>	<i>7</i>
<i>Other Project-related Work</i>	<i>8</i>
<i>Advisory Board.....</i>	<i>8</i>
<i>Project Meetings.....</i>	<i>9</i>
4. DISSEMINATION ACTIVITIES.....	10
<i>Online.....</i>	<i>10</i>
<i>EPO Collaboration</i>	<i>10</i>
<i>Publications</i>	<i>11</i>
<i>Conferences and Presentations</i>	<i>11</i>
CONTACT	12

1. Executive Summary

The PLUTO project (Patent Language Translations Online) aims to provide a rapid solution for the online retrieval and translation of patent documents through the integration of a number of existing state-of-the-art components provided by the partners in our dynamic industry-academia consortium.

It is intended that the provision of this service will have a significant positive impact on the intellectual property (IP) related activities of a number of parties; individual inventors and Small- and Medium-size Enterprises (SMEs) with potentially limited resources, as well as multinational corporations and other entities, will be afforded the ability to carry out prior art searches in a much more effective and efficient manner through a reduction in time and translations costs when using PLUTO services.

Additionally, PLUTO aims to advance the state-of-the-art in domain-focused machine translation (MT) through its integration with translation memories (TM). A number of novel approaches to cross-lingual information retrieval will also be employed during the construction of a multilingual searchable database of patents.

Since the project kick-off meeting in May 2010, PLUTO has seen rapid progress toward its intended goals due to a combination of consortium partners' existing working relationships, the strength of products and systems upon which the project builds, and close alignment with pilot initiatives for Machine Translation at the European Patent Office.

All internal milestones of the project have been met successfully, with data being provided for TM and MT development, documented integration contracts detailing APIs and data requirements among the core PLUTO system components, as well as development of core components (search engine, TM and MT) progressing around shared data and languages for the 12-month prototype system (due April 2011).

Dissemination activities are already well underway with the basic infrastructure of project logo, web site, fact sheet, brochure, presentation etc. being complemented by partner dissemination activities at relevant international conferences and events.

In addition to planned project milestones and deliverables, a major highlight of our achievements during the first calendar year relates to the PLUTO project's engagement on a pilot Machine Translation project for English–Portuguese with the European Patent Office. We were able to successfully develop and deploy a machine translation system with EPO data, pass stringent engineering testing and quality review by the national patent office, and achieve the EPO project officer's recommendation that the system be accepted for use at the EPO's Espacenet service. The system has been in live production use at Espacenet since the beginning of October 2010.

2. Project Overview

A core aspect of the European Commission's (EC) commitment to language diversification is the provision of multilingual access to intellectual property information, namely patents. This will afford inventors in Europe better access to technical information on patents in their native language and foster innovation and growth. Central to such a provision is the availability of high-quality search and translation technologies capable of dealing with the volume and language diversity of large collections of patent data. MT software must also be adapted to handle the specific language found in patent documents.

In order to adequately facilitate this, the PLUTO project aims to develop an online framework – through the integration of a number of mature software components – whereby users can exploit state-of-the-art tools to search large patent repositories and automatically translate the retrieved documents.

As well as directly addressing the translation needs of the EC, PLUTO serves a more general purpose when it comes to IP related activities. SMEs and individual inventors can encounter difficulties when entering a new market due to the high costs related to prior art search and translation. Often, making such a leap constitutes a large risk for these entities. PLUTO aims to reduce the risk by providing an integrated, online translation tool where clients – in the form of several human experts (technical, legal, consultants) – can take advantage of existing web-content and state-of-the-art MT technologies and information retrieval (IR) tools to collaboratively search for, retrieve and translate patents in a fast, cost-effective manner.

In doing this, PLUTO will also advance the state-of-the-art in MT through novel approaches to integration with TM and domain adaptation techniques aimed at dealing with the specific characteristics of patent documents (legalese, technical terminology and long sentences). Furthermore, a number of innovative approaches to cross-lingual information retrieval will allow for more efficient patent searching.



Members of the PLUTO consortium with PLUTO Project Officer at project kick-off meeting in Vienna

Consortium

The PLUTO consortium comprises both academic and industrial partners from across Europe with high levels of expertise in their respective areas. Each of our members brings unique knowledge and competency to PLUTO and affords the project a great opportunity for success.

The Centre for Next Generation Localisation at Dublin City University brings to the target platform a state-of-the-art machine translation engine, MaTrEx, which exploits hybrid statistical, example-based and hierarchical techniques and has demonstrated high-quality translation performance in a number of recent evaluation campaigns. In addition to significant commercial experience and expertise, ETeam AB contributes a comprehensive translation software environment to the project, including server-based, multi-layered, multi-domain translation memory technology.

Information retrieval expertise is provided by the Information Retrieval Facility which also provides access to its data on patent search use-cases and a large scale, multilingual patent repository. PLUTO will also exploit the use-case holistic machine translation expertise of Cross Language, who have significant experience in the evaluation of machine translation quality, while the Dutch Patent User Information Group, WON, will be directly involved in all phases of development, providing valuable feedback on the usability and appropriateness of the service.



CNGL partners participating in the kick-off via video conferencing. Travel plans were curtailed due to the Icelandic ash cloud.

3. Progress to Date

Achievements

Despite the project being in its relative infancy, significant progress has been made towards meeting a number of project milestones and goals. Some key achievements to date include:

- Deployment of an online production-level MT service for English–Portuguese at the European Patent Office;
- Development of a first prototype for the PLUTO web service and patent search engine;
- Specification of an integration framework for the individual software components involved in the overall service.

EPO Collaboration

Following a PLUTO project briefing to the European Patent Office Machine Translation project team at Den Haag in June, a request was received by the PLUTO project coordinator to assist with a pilot project in MT for the English–Portuguese language pair. Having consulted with the project consortium, and given how closely aligned the pilot project was with the PLUTO project goals, it was agreed that this project should be undertaken.



This collaboration has provided a number of distinct technical benefits for the consortium including the receipt a large quantity of patent data, some of it of a very high quality. In building MT engines for the EPO and developing a full-scale production-level web service through which the translation service was delivered, a number of specific project requirements were also satisfied ahead of time.

The preparation of hosting arrangements at DCU for the MT web service aided in the development of a number of APIs and interfaces which now serve as core elements of the PLUTO integration framework. Furthermore, upon implementation of the translation service, we came through a thorough evaluation, provided by the EPO, from both a linguistic perspective in terms of the quality of the translations being produced, as well as from an engineering standpoint, in terms of the speed and robustness of the web service.

Other Project-related Work

In addition to the aforementioned achievements and outputs, substantial work was carried out in line with the predefined work plan of the project.

- Patent data was made available to the consortium in a number of languages from two sources: the IRF's MAREC corpus and the European Patent Office (EPO).
- Project partners have defined the XML schema via which documents will be exchanged between the search engine, the translation memory, and the machine translation (MT) system.
- Translation memory resources have been generated from the data supplied from the IRF and the EPO and added to the repository at ESTeam.
- As well as the MT engines for English–Portuguese, further engines for English–French have been developed including investigation into a number of approaches for domain adaptation according to the International Patent Classification system.



PLUTO poster presentation at EAMT 2010

Advisory Board

The consortium profiled the type of individuals it would be most beneficial to have on our advisory board, namely: an MT expert, an IR expert, an IP expert and someone from a user group e.g. the Patent Information User Group. To this end, we can confirm the following individuals as members of our advisory board:

- Dr. Fred Hollowood, Symantec
- Dr. Gregory Grefenstette, Exalead
- Viggo Hansen

Project Meetings

A number of consortium meetings have been held over the course of the first six months of the project. They are detailed below.

- **Project Kick-off Meeting** (May 5/6th 2010, Vienna, Austria)

In addition to PLUTO Project consortium members, this meeting was attended by the PLUTO project officer and representatives from the EPO. Due to the grounding of flights from Ireland due to the “Ash Cloud”, representatives from DCU attended the meeting via video conference.

- **Technical Management Board Meeting** (June 21/22nd 2010, Den Haag, the Netherlands)

Member of the technical teams from all core partners met in Den Haag on June 22nd. This was preceded by a meeting of the PLUTO consortium with representatives of the Machine Translation project at the European Patent Office on June 21st. At this meeting an overview of the PLUTO project was given to the EPO and the EPO in turn provided an overview of Machine Translation initiatives underway at EPO. This set the foundations for the collaboration with the EPO.

- **Project General Assembly** (October 18th/19th 2010, Dublin, Ireland)

A full two-day general assembly was held in Dublin to review progress in all work packages over the course of the first six months of the project. Six month project deliverables were also reviewed and discussed prior to submission. Finally a more in-depth technical working session was held to finalise the integration framework for components of the PLUTO system.



Members of the PLUTO consortium during a General Assembly at Dublin City University.

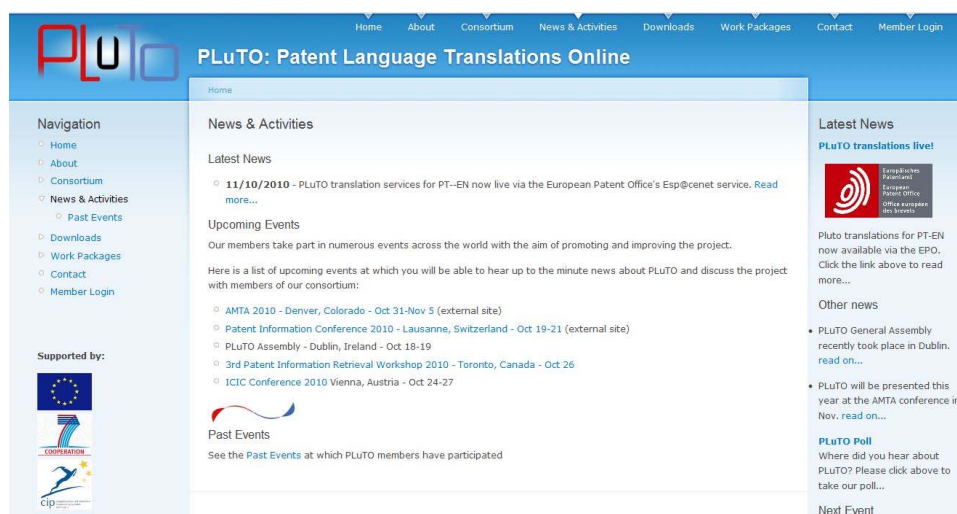
4. Dissemination Activities

A dissemination plan for the project was developed as a deliverable over the course of the first six months of the project. Our strategy is to promote the project and its output through a number of channels, including:

- The project website;
- Social media outlets;
- At conferences through publications, presentations and demos/showcases;
- Via press releases and other communication materials.

Online

The PLUTO website (<http://www.pluto-patenttranslation.eu>) went live on July 16th 2010. The site provides access to consortium information, updates on news items and links to downloadable content, e.g. publications, fact sheets, presentations, etc. The project also has a Facebook page and a presence on Twitter via the hashtag #PLUTO_EU.



A screen cap of the PLUTO website – <http://www.pluto-patenttranslation.eu>

EPO Collaboration

We have been able to capitalise on the collaboration with the EPO as a means to further promote the project and to satisfy some dissemination requirements. Firstly, the technical aspects of the work form the core of the publication accepted at the AMTA conference (see Publications below). Additionally, the service which the consortium provides to the EPO has been publicised on both the website of the project as well as the website of the EPO (<http://www.espacenet.com/>) and can be referred to in future dissemination materials. We also intend to source

testimonials from the EPO and users of the translation engine to promote the project from a commercial stand point.

New in esp@cenet

As and from today, the online machine translation tool on *esp@cenet* allows **Portuguese** language abstracts, descriptions and claims to be translated into English and English language abstracts, descriptions and claims to be translated into Portuguese.

News item on <http://www.espacenet.com> on the PLUTO provided PT translation service

Publications

We have had one paper accepted to the commercial user track of the AMTA conference (American Machine Translation Association). The paper is publicly available for download from the project website and its reference is given below:

- **John Tinsley, Andy Way and Páraic Sheridan. 2010. *PLuTO: MT for Online Patent Translation*. In Proceedings of AMTA 2010: The Ninth Conference of the Association for Machine Translation in the Americas, Denver, CO.**

Conferences and Presentations

- **EAMT 2010 (May 26/27th, St-Raphael, France)**

PLUTO was presented at an FP7 project showcase held as part of the conference of the European Association for Machine Translation. The conference was attended by PLUTO representatives from DCU and Cross Language. The project was introduced during a short oral presentation and as a poster during a dedicated session. Additionally, the project fact sheet was made available as a handout in the main auditorium of the conference.

- **IRF Symposium 2010 (June 1–4th, Vienna, Austria)**

PLUTO was exhibited in the main hall at the IRF Symposium by means of a large backlit column describing IRF activities. PLUTO was also referred to specifically in the programme for the meeting. The symposium was attended by a number of members from the IRF.

- **SIGIR 2010 (July 19–23rd, Geneva, Switzerland)**

PLUTO was presented as a poster at the IRF booth in the exhibition hall at the conference of the ACM Special Interest Group on Information Retrieval. Attended by members of the IRF, the project was the topic of much interesting discussion over the course of the 5 days.

- **AMTA 2010** (November 1–4th, Denver, USA)

PLUTO was presented during a session of the commercial user track of the AMTA conference. This 30 minute talk was based on the aforementioned publication and the conference will be attended by members from DCU.

- **IPWare Summit 2010/2011**

PLUTO was invited to give an hour long presentation at the IPWare Summit in 2010. Unfortunately this conference was postponed. However, the consortium has been made aware that the invitation still remains for the next edition of the conference in 2011.



PLUTO presentation at the Commercial User Track of the AMTA 2010 conference in Denver.

Contact

If you have any questions or comments related to the information in this document please contact one of the following people:

Name:	Dr. Páraic Sheridan (PI)	Dr. John Tinsley (Coordinator)
Phone:	+353-1-7006706	+353-1-7006916
Email:	psheridan@computing.dcu.ie	jtinsley@computing.dcu.ie