PROGRESS REPORT

§3 Project Progress

Grant Agreement number: 250416

Project acronym: PLuTO

Project title: Patent Language Translations Online

Project type: Pilot B

Periodic report: 1st

Period covered: **from** 01/04/2010 **to** 31/03/2011

Project coordinator name, title and organisation: Dr. Páraic Sheridan, DCU

Tel: +353-1-7006706

Fax: +353-1-7006702

E-mail: psheridan@computing.dcu.ie

Project website address: http://www.pluto-patenttranslation.eu

Authors:

John Tinsley (DCU)

Executive Abstract

In this report, we describe the work carried out over the first reporting period of the PLuTO project. We present the overall objectives of the project, with specific focus on those objectives falling due in this period. For each work package, the main goals, tasks completed, milestones and deliverables achieved, and future plans are presented. In a separate section, we give an overview of the activities of the project coordinator in this period, outlining details on collaborations, project internal activities, and issues which arose. We conclude by outlining plans for the coming period.

The main highlight of the project to date has been the consolidation of development and efforts in a number of key areas – data provision, machine translation (MT) engines building, translation memory resource creation, and web application development – into a fully-functional integrated web-based prototype. The prototype exists as a demonstrable output of the project affords the consortium the opportunity to step up user feedback, dissemination, and exploitation activities.

A further highlight of the period was the collaboration between the consortium and the European Patent Office (EPO). A bespoke English—Portuguese MT engine for patent translation was developed for, and used by, the EPO in their Espacenet service for a period of 6 months. MT was provided as a robust, production-grade web service hosted at Dublin City University and provided the consortium with a significant stepping stone in the development of the integrated prototype system.

Plans for the second period are heavily driven by the feedback received to date from the project advisory board and user group, and feedback expected from an evaluation survey of a wider user base. Development will continue on improvements to translation quality, expansion of the coverage of the service, and implementation of value-adding features to the prototype system as we move towards a more production oriented environment.

(Disclaimer)

This report is still in draft format. While the majority of the sections are complete, there are a few details which remain incomplete at this time. The most significant sections yet to be filled are those containing details on the resources (person months) expended in each work package, and the concluding summary.

1 Project Objectives for the Period

The overall aim of PLuTO is to a develop a rapid solution for online patent translation services through the integration of a number of existing components including the MaTrEx Machine Translation (MT) engine of DCU, ESTeam's Translation Memory (TM) technology and the search tools and patent repository of the IRF. Iterative improvements will be made to the integrated platform over the course of the project guided by the outcome of evaluations carried out by Cross Language and feedback from the WON user group.

In terms of non-technical objectives, dissemination and exploitation activities are necessary to ensure that the activities outputs of the project are made known to the relevant communities to ensure maximum impact and long-term sustainability.

In order to achieve these aims in the context of the first year of the project there are a number of intermediate objectives to meet. The primary goal is the development of a first online demonstration prototype of the PLuTO target platform as outlined in Deliverable 6.1. Development of this prototype requires the completion of a number of sub-tasks which essentially involves the development of the individual components which together comprise the prototype. As such a software integration task is non-trivial, an integration framework must be designed in order for all the components to communicate and function efficiently as one.

In addition, initial dissemination activities will be carried out to introduce the project to relevant audiences and an analysis of the commercial landscape into which the fruits of the project will ultimately be released will also take place.

Specifically, the objectives of the consortium for the period described in this report amount to:

- The development of MT engines for 2 language pairs (EN, FR, PT);
- The creation of TM resources for the above languages;
- The provision of relevant patent corpora to support the aforementioned tasks;
- The production of a web application and interface through which users can access the selection engine, translation services and other functionally of the platform;
- An initial evaluation of all relevant components including translation quality and adequacy, relevance of search results, and usability of the service;
- The implementation of a number of management tools, e.g. reporting procedures and quality management processes, to facilitate coordination of the project;
- The preparation of a dissemination plan including development of a project website and presentations on the project at a number of relevant events.
- The preparation of a feasibility report on the founding of a joint SME between project partners based on the technical outputs.

The measurable objectives in this period will serve as a benchmark by which the objectives of subsequent periods will be evaluated

2 Work Progress and Achievements during the Period

In this section, we describe each work package in detail outlining the global objectives, progress made during the period, followed by specific details on the individual tasks set out in the Description of Work. Work package 1 – Management – is excluded here as it is treated as a standalone topic in section 4.

2.1 WP2 Data Acquisition, Selection and Integration

2.1.1 Objectives

The global objectives of this work package are to ensure the constant availability of patent data to the consortium for the purpose of training MT engines and producing TM resources and to provide the information retrieval specific tools to allow for selections (search) of data. Deliverables and milestones falling due in the period are shown in Table 1.

Mi2.1	(EN—PT) patent data available	☑
Mi2.2	(EN, PT, FR) patent data available	☑
Mi2.7	Metadata annotation sets and guidelines available	☑
Mi2.8	Final metadata annotation sets and baseline seclection engine available	☑
D2.1	Data Corpora and Standards, v1	☑

Table 1 Milestones and deliverables due between M1 and M12

2.1.2 Progress Highlights

As stated, the key objective of this work package is to provide data across a number of language pairs to be used for search, MT training, and TM building. To this end, data has been provided for En—Fr and En—Pt from two distinct sources: the IRF's MAREC corpus and the EPO.

Additionally, a set of metadata definitions has been agreed upon to define the interchange format between the various components of the system. This includes standard XML mark-up as well as supplementary descriptive information on the data.

Finally, a baseline search engine has been prepared which indexes the data found in the MAREC corpus. This is exposed to the integrated system via a web service.

2.1.3 Tasks

T2.1 Meta-data definition

A metadata definition has been agreed across the partners as the format of the data is integral to a number of the key components, including the search engine and the MT engine (input/output formats), and for MT/TM integration.

There are a number of important fields in the mark-up which contain information important to different system components. For example:

- *family-id* – this allows for the identification of related documents which can be useful for search and evaluation of search;
- *lang* – this tells us what language a patent document is which supports the search and translation engines;
- *kind* – this indicates which IPC domain the patent belongs which allows us to exploit domain-specific features of the various components;
- *alignment* – this fields appears in MT output and describes the segments used by the engine to compose the final translation. This is used during MT/TM integration.

More information on the mark-up and metadata is provided in Deliverable 2.1 Data Corpora and Standards and Mi6.1 Integration Contracts.

T2.2 Selection Engine

The baseline search (selection) engine for PLuTO has been developed based on the Apache SOLR engine. It is currently exposed via a secure web service and is capable of answering search queries and retrieving full text documents from the MAREC collection.[1]

To date, the baseline engine has not been adapted to the peculiarities of patents as the focus for year one was on integrating the engine with the other PLuTO components. The engine allows for the incorporation of a query translation service which currently can send calls to the Google translation API. This will serve as the baseline against which any query translation service produced in work package 5 will be compared.

The web service through which the search engine is exposed is based on a RESTful architecture and is fully integrated within the PLuTO framework. Full details on the search engine are provided in the report on Milestone 2.8.

T2.3 Data Acquisition

The consortium has exploited patent data in two language pairs over the course of the first year of the project: English—French and English—Portuguese. This data has come from two sources.

Firstly, the En—Fr data was provided as part of the MAREC collection. This is a unified collection of patent documents provided by the IRF that is represented in XML format following a standard document type definition (DTD). In addition to this data, through collaboration with the EPO a set of English and Portuguese patent documents were provided to the consortium along with a number of other resources, e.g. a translation memory and bilingual dictionaries.

Plans are currently in place to acquire additional data for English—German and English—Dutch for year two. En—De data is already available as part of the MAREC collection, while a decision was made to treat En—Nl at this stage to best exploit the expertise of our partners

[1] This does not include the Portuguese data provided by the EPO. We plan to index this data in order to add it to the search engine.

WON (Dutch Patent User Group) and Cross Language, the majority of whose employees are native Dutch speakers. The IRF are currently undertaking work to source Dutch patent data.

Further details on the content of the respective corpora, their metadata, and the standard to which they comply can be found in Deliverable 2.1.

2.1.4 Use of Resources

Beneficiary	PMs Actual (Year 1)	PMs Planned (3 years)
DCU		
ESTeam		
IRF		
Cross Language		
WON		
Total		

2.1.5 Summary

We have presented the patent corpora which have been compiled to date for exploitation in the major technical components of the PLuTO system. Metadata definitions for XML mark-up and exchange formats have been defined across partners and Plans have also been implemented for acquiring data for year two.

Additionally, a baseline search engine has been deployed and integrated with the other components via a web service.

2.2 WP3 Web Application and User Interface

2.2.1 Objectives

This main goal of this work package is to design, implement and provide a collaborative online environment to the end-user of the PLuTO service and to connect the patent data, selection engine, and integrated translation services back-ends. Deliverables and milestones falling due in the period are shown in Table 2.

Mi3.1	First web application and user interface	☑
D3.1	First web application and user interface	☑

Table 2 Milestones and deliverables due between M1 and M12

2.2.2 Progress Highlights

The web application interface, which serves as the first point of contact a user will have with the PLuTO system as well as the back-end through which all technical components communicate, has been developed. This provides a single point of access to the core search and translation components of the entire architecture.

In addition to the user interface, the initial administrative interface has been designed and implemented. This allows for a membership and authentication service whereby users can

create accounts to use the service. By means of a back-end database layers, user account details are stored along with the capability for logging and gathering of statistics on sessions.

The application is the final piece of the puzzle which gives rise to the first integrated prototype, described further in section WP6 System Integration2.5. With the web interface in place, we can allow access to a first wave of users who can provide us valuable feedback on usability, existing features and suggested features.

2.2.3 Tasks

T3.1 Data Layer

The data access layer (DAL), or data layer, concerns all instances in which data is read or written when using the PLuTO service. It is the component in the web application that connects the user interface to the various data repositories, e.g. patent documents and statistics/logs.

The DAL is implemented as the lowest layer in the overall system architecture, which is modelled on both service-oriented and N-tier architectures. This allows for extensibility and reusability of the various system components.

The database used to store information is a Microsoft SQL Server 2008 R2. A full diagrammatic view of the database schema can be seen in section 4.5 of Deliverable 3.1. and in the various appendices to the deliverable document.

T3.2 User Interface

The user interface concerns the means by which the end-user will interact with the system; essentially the web-based GUI. The user is first presented with a login screen. Once logged in, the user can perform patent search by title and/or publication date. Following this, the user can download the patent document, view bibliographic data, claims, description, etc. At this point, the user has the option of sending sections of the patent to be translated. User authentication is secured by 64bit encryption in the first prototype. This will be increased in subsequent releases.

The service also includes administrative and business interfaces. These clients provide facilities for the management of resources such as user accounts. It also provides logs and statistics on usage and provides access to an external help desk module where support personnel can provide assistance to users.

T3.3 Application Interface

The application interface addresses the definition of the web services through which various components, such as the search and translation engines, can interact with the web application.

The service layer has been designed to handle various web service protocols, such as SOAP and RESTful, while the services themselves support multiple protocols such as: HTTP, TCP, Names Pipes, and MSMQ.

Full details on the application interface, as well as the overall system architecture and layers can be found in Deliverable 3.1.

2.2.4 Use of Resources

Beneficiary	PMs Actual (Year 1)	PMs Planned (3 years)
DCU		
ESTeam		
IRF		
Cross Language		
WON		
Total		

2.2.5 Summary

The core architecture behind the PLuTO web application has been designed and implemented. This includes the first instance of the user interface to the search and translation features. User and administrative authentication features have been enabled along with the capacity to manage accounts, track and log usage, and view statistics.

Additionally, the database layer has been designed which provides efficient access via the web application to the patent and other data repositories.

2.3 WP4 Translation Memory

2.3.1 Objectives

The key aim of this work package is to create translation memory (TM) resources from the data provided in work package 2. This involves pre-processing of the raw data, structuring of the data based on the IPC system and alignment of multilingual segments. These resources are then to be exposed in a database to other components (web application, MT engine) via web services defined in work package 6 and implemented in work package 3. Deliverables and milestones falling due in the period are shown in Table 3 Milestones and deliverables due between M1 and M12

Mi4.1	TM resources for EN-PT	☑
Mi4.2	TM resources for EN-FR	☑
D4.1	TM resources for 2 language pairs	☑

Table 3 Milestones and deliverables due between M1 and M12

2.3.2 Progress Highlights

To date, translation memory (TM) resources have been created for the two language pairs addressed in this period: English—French and English—Portuguese. This process involved cleaning and alignment of the data provided as described in section 2.1 and the subsequent loading of the data into a database according to the IPC structure.

Additionally, common TM entries in the pivot language, English, have been identified across the two TMs which have allowed for the creation of a French—Portuguese TM. This technique allows us to significantly expand the language coverage of the translation service over the course of the project.

2.3.3 Tasks

T4.1 Data Management

Relevant data for both language pairs was extracted from the available corpora and a number of pre-processing steps were applied, including language identification, filtering of noisy data, and segmentation.

T4.2 Structuring TM domains according to patent data domains

For the TM resources, data is structured according to the IPC system. Given the sparseness of our available data relative to the extremely fine-grained hierarchical nature of the IPC system, the TMs were structured according to only the top three levels.

T4.3 Alignment

The TMs were aligned at a number of levels using the ESTeam Translation Management (TraM) interface. Following alignment at the file level, data level alignment is carried out at paragraph, sentence, segment, and sub-segment level.

T4.4 Data loading and quality control

Following manual checking of the alignments, the individual bilingual entries are marked up according to the metadata specifications described in sections 2.1 and 2.5 and loaded into the TM module.

2.3.4 Use of Resources

Beneficiary	PMs Actual (Year 1)	PMs Planned (3 years)
DCU		
ESTeam		
IRF		
Cross Language		
WON		
Total		

2.3.5 Summary

Translation Memory resources have been created for the two language pairs addressed by the project in this period, as well as for French—Portuguese using pivoting techniques. Data has been cleaned, aligned, and loaded into the database ready for integration with the MT engine. Full details of this process along with statistics on the data are provided in Deliverable 4.1.

2.4 WP5 Machine Translation

2.4.1 Objectives

The principal goal of the Machine Translation (MT) work package is to build MT engines for the language pairs being addressed in the project using the MaTrEx system. In order to achieve optimal performance, this will require the adaptation of the MT technology to the patent domain. Deliverables and milestones falling due in the period are shown in Table 5.

Mi5.1	MT engine for EN-PT	☑
Mi5.2	MT engine for EN-FR	☑
D5.1	MT engine for 2 language pairs	☑

Table 4 Milestones and deliverables due between M1 and M12

2.4.2 Progress Highlights

In the context of this work package, the main achievement has been the development of Machine Translation engines for two language pairs: English—Portuguese and English—French.

The En—Pt engine was developed in the first six months of the project in conjunction with the requirements of the EPO collaboration (described further in section 4.1.1). Data for this language pair, described in detail in Deliverable 2.1, was provided by the EPO. The system was fully evaluated on two separate occasions by the EPO and the Portuguese National Patent Office and deemed to be of sufficient quality to use for their online patent search and translation service, Espacenet,[2] between September 2010 and March 2011. This also resulted in the development of a production-level web service for MT which has subsequently been used in the first integrated prototype.

The En—Fr engine was the first example of a system built using parallel data provided explicitly for use within the consortium. Relevant (parallel and comparable) data was extracted from the MAREC patent collection described in section 2.1 and pre-processed for MT training. This task represented the first significant interaction between work packages 2 and 5.

2.4.3 Tasks

T5.1 Adapting existing MT technology to the patent domain

Patent translation is a unique task given the nature of the language found in patent documents. Across the abstracts, claims, and descriptions there exists a mix of legalese, scientific writing, and specific terminology related to the topic of the patent. Given this, the task of building MT systems is not as straightforward as collected large collections of patent data and training a single system. Instead, we must consider how to adapt the systems to best handle the specific type of patent being translated.

We carried out a number of experiments for both En—Pt and En—Fr in which combinations of in-domain and general domain data (according to the IPC system) were used to train the various models of the MT systems. Full details on the experimental set up can be found in Deliverable 5.1.

Our main findings from these experiments were that there are some benefits to be had from exploiting out-of-domain/general data when translating documents in a specific domain. In many cases the best translations were achieved using an in-domain translation model with a general language model. However, translation performance will always significantly affected

[2] http://www.espacenet.com

by the distribution of data; that is to say how much data is available for each domain. For instance, we will obviously have better translations if we use a general TT when there are only, say, 5000 sentence pairs available for training in a specific domain. We also identified certain domains which do not benefit as much from out-of-domain data, e.g. domain C (Chemistry; Metallurgy), which can be attributed to the type of language found in these patents; chemical names, formulae, compounds, etc.

In addition to domain adaptation based on the IPC system, we also carried out preliminary investigations as to the feasibility of employing separate MT systems for the different sections of a patent document, notably abstracts and claims.[3] Initial findings suggest there is not much benefit to be had from taking this approach but we may revisit this question depending on the availability of data for future language pairs.

T5.2 Integration between SMT/EBMT and RBMT

In the current implementation of our MT architecture, we employ one EBMT specific technique and one set of rules to adapt our systems to some particular characteristics of patent documents.

Firstly, we use a number of rules and other mark-up to ensure that references to figures and tables, other patents, and numerical and/or alphabetised lists found in patent documents are handled robustly. The rules we employ first identify such items in the input and removes them. They are then reinserted at the appropriate position in the translation output based on alignment information we cache during the decoding processes.

Secondly, one factor which makes patents particularly difficult to translate in the length of the individual sentences. This is exacerbated by the fact that each 'claim' must be expressed as a single sentence. It is well established that MT quality suffers the longer and more complex each input segment is. To overcome this, we exploit an EBMT technique based on the "marker hypothesis" which splits input sentences into smaller, more translatable, segments based on a set of closed-class (or "marker") words. These segments are then translated independently and recombined into a single sentence post-translation. As well as improving translation quality, this method also allows us to increase the efficiency of our translation pipeline through improved parallelisation. This is described in greater in detail in Deliverable 5.1

T5.3 Integration between MT and TM

The first step in providing for MT and TM integration was to deploy the MT system as a web service. The MT service is hosted at DCU while ESTeam hosts the TM component. Communication takes via an API agreed between the partners and specified in the Integration Contracts document (Mi6.1).

Patent documents are sent for translation in XML format via a URL using the XML-RPC protocol. The URL contains information such as the translation direction while the XML mark-up in the document contains information of the IPC code document section (abstract, etc.) should we require it. The XML is then processed and the data is sent to our translation servers in the appropriate format.

[3] Abstracts and claims are the sections of patent documents for which we had parallel or comparable data. There was minimal data of this kind available of descriptions.

Following translation, the XML document is reassembled in the target language. Additional mark-up is added to the output which provides alignment information on how the MT system assembled the translation. This is then exploited by the TM component to allow for integration of the TM options with MT output to produce optimal translation output. The methodology for integration is described below in section 2.5 and in detail in Deliverable 6.1

In order for the translation web service to return translations as quickly as possible, the translation server distributes tasks multiple workers/cores, translating up to 8 sentences simultaneously. This architecture is based on the multiple producers/consumers pattern and ensures that all translation jobs receive the same share of computational resources regardless of their size or when they were submitted. The complete server architecture is described in detail in Deliverable 5.1

2.4.4 Use of Resources

Beneficiary	PMs Actual (Year 1)	PMs Planned (3 years)
DCU		
ESTeam		
IRF		
Cross Language		
WON		
Total		

2.4.5 Summary

We have highlighted the areas in which we have made significant progress over the course of the first period. These include the deployment of MT system for two language pairs via an efficient web service which allows for integration into the prototype system and with the TM module of ESTeam. We have also demonstrated efforts made towards domain adaptation of MT for patents as well as integration of our core SMT approach with techniques from rule-based and example-based paradigms.

The experiments summarised in this section are described in detail in Deliverable 5.1 and have resulted in two peer reviewing publications: Tinsley et al. (2010); Ceausu et al. (2011).

2.5 WP6 System Integration

2.5.1 Objectives

The System Integration work package will provide the technical framework in which the various software components in PLuTO. The main outcome of the work carried out here will be the delivery of the first PLuTO system prototype that integrates MT, TM, and search functionality together through web services and is accessible to the end-user via the web application of work package 3. Deliverables and milestones falling due in the period are shown in Table 5.

Mi6.1	Integration contracts completed	☑
Mi6.2	Integration prototype available	☑

D6.1	First prototype	☑

Table 5 Milestones and deliverables due between M1 and M12

2.5.2 Progress Highlights

The main highlight has been the production of the first integrated prototype of the PLuTO system. Based on the survey of the various components' architectures as part of Milestone 6.1 Integration Contracts, a web service and data exchange formats were defined in the first six months of the project. The following six months were spent integrating the components. The DCU MT engines and the IRF retrieval system were exposed via web services. ESTeam developed the web interface and integration modules in work package 3 which linked these services together along with the TM resources. This integrated prototype is now available as a functional demo system and will be presented at the first review.

2.5.3 Tasks

T6.1 Integration Requirements Analysis

The aim of this task was to determine the exact technical requirements for integrating three components – search, MT, and TM – which were developed completely independently of each other. This involved two steps: a requirements analysis by carrying out a survey of each system, determining its architecture, dependents, I/O formats; and, based on the analysis, drawing up the integration contracts which define the processes through which the components will communicate.

- Requirements Analysis

Following an analysis of the various systems, it was determined that the optimal solution for the first prototype was to have the individual components hosted locally at their respective partners' sites, i.e. MT systems hosted in DCU, etc., and communicate via web services. This was due to the significantly different architectures of the systems. The ESTeam software is solely Windows based, while the DCU software was developed in UNIX and has not been tested on other platforms. While this is not the best solution going forward in terms of developing an efficient production environment, it was decided it would be better to first get the prototypical integration implemented remotely for the sake of meeting our time constraints and subsequently moving the systems to a single location by the end of the project.

- Integration Contracts

Once the requirements for integration were understood, the integration contracts were drawn up to provide the specifications for the prototype. Decisions made here were heavily inspired by the projects' collaborative efforts with the EPO in which the majority of the software developed was reusable for project specific purposes. It was decided that the web services would be implemented as a Java-based RESTful service whereby requests could be sent to the various components via a URL. The XML exchange format was defined for sending and receiving documents and error handling and authentication methods were defined. This document can be seen in the report on Milestone 6.1.

T6.3 Integration Prototype Generation

Given the specification of the requirements in the integration contracts, the first prototype was implemented accordingly. The prototype is web-based and the first point of contact for the end-user is a login page. After logging in (as a user or an administrator), the user is

presented with the patent search engine and can perform multi-lingual search in the 3 languages addressed by the project so far. Upon selecting a document from the list of search results, the user has the option of sending it for translation. At this point, the integrated MT/TM system is invoked and the translated document is returned. The translation service can also be accessed directly through a distinct user interface.

The prototype system is described and reviewed in Deliverable 6.1 and will be demonstrated at the year one review.

2.5.4 Use of Resources

Beneficiary	PMs Actual (Year 1)	PMs Planned (3 years)
DCU		
ESTeam		
IRF		
Cross Language		
WON		
Total		

2.5.5 Summary

The first integrated prototype of the PLuTO system has been developed. It is a web-based service which, when accessed through a user interface, provides a fully functional patent search and translation system.

In the current implementation, the various components are located remotely from one another and communicate via predefined web services. The services will be evaluated by WON according to the procedures described in section 2.6 and all recommendations will be taken into account and used as a basis from implementation of subsequent versions of the prototype.

2.6 WP7 Evaluation and Quality Assurance

2.6.1 Objectives

The ultimate goal of the Evaluation work package is to ensure that the final PLuTO integrated platform meets the needs of patent searchers, patent user groups, and other potential users. In order to ensure the required standards are met, the individual components of the system must undergo a thorough evaluation and quality assurance process throughout the duration of the project. This is carried out not only through qualitative evaluations of document retrieval and translation quality, but also by engaging WON on the overall usability of the system taking the individual components into account. Deliverables and milestones falling due in the period are shown in Table 6.

Mi7.1	Survey structure and content available	☑
Mi7.4	Integrated cross-lingual evaluation framework – report	☑

D7.1	Survey structure	☑

Table 6 Milestones and deliverables due between M1 and M12

2.6.2 Progress Highlights

The main highlights to date in the context of this work package have been the evaluations and analyses of our MT systems. The En—Pt system was twice evaluated by project external bodies: once by the Portuguese National Patent Office, and once by the EPO. These evaluations provided not only a sense of the general quality of the translation output but also important details on the usability of the translations regardless of linguistic/grammatical correctness, categorisation of errors and the productivity of the system in terms of speed.

In addition to the external evaluation, we carried out a systematic internal evaluation of our En—Pt and En—Fr systems in both translation directions. This included an automatic and systematic manual evaluation as well as comparative analysis against the Google and Systran systems.

An evaluation survey was prepared by Cross Language and DCU in conjunction with WON as a requirement of Deliverable 7.1. Members of the consortium met with the WON PLuTO working group following the April AGM to discuss the contents of the survey and the expectations of the group in relation to it.

Finally, a report was prepared on the framework for evaluation of information retrieval quality in the context of the PLuTO engine.

2.6.3 Tasks

T7.1 Usability and utility to patent searchers

The specific objective of usability evaluation is to ensure that the overall system meets the needs of the patent searchers and patent users groups. MT usability evaluation is mainly user centred and takes into account use cases of translated text, which goes beyond the classical approach in MT evaluation. This includes a simulation of typical user tasks with translated text.

Through consultation with the PLuTO working group at WON, as well as members of our advisory board, a survey was prepared to collect valuable information from patent users. The survey will provide us with both objective and subjective feedback on the usability and adequacy of our machine translations in addressing the needs of patent users. Additionally, a number of direct questions in the survey will give us a better idea of where to focus our efforts in the project going forward, e.g. increase resources into translation over search, what languages to treat, and features to include in our interfaces.

The survey is presented in full as Deliverable 7.1., and a full report on the results and findings from the survey will be released in M18 (Oct 2011) as Deliverable 7.2.

Significant insight into the usability of our En—Pt system was also derived from the evaluations by the EPO. We received evaluation summaries on a document by document basis with information on the overall impression of the translation as well as specific errors. Combining all the reports together, we were able to garner a high level picture of areas in which the MT system was consistently erring, and focus improvements on these. This led to us making various improvements to our pre- and post-processing stages to allow for

improved formatting of our output, including maintaining the integrity of chemical symbols, references, lists and headings.

T7.2 Retrieval Evaluation

The retrieval evaluation framework aims to provide us with a continuous feed of measurements of both efficiency and effectiveness of the retrieval process. Efficiency measures relate to the computational resources used and the time taken to return search results. This is done using monitoring facilities built into the Solr engine.

Measuring of the effectiveness of the retrieval engine is carried out using both automatic and manual methods. Based on the experience gained by the IRF during their organisation of the CLEF-IP and TREC-CHEM evaluation campaigns, a number of measures will be used to assess the adequacy of retrieval results. Automatic measures such as MAP, NDCG, P@10, R@100, and PRES will be applied via the web service through which the engine is exposed. Future plans for a user-centric evaluation include the addition of features to the web service through which users, including WON during specified evaluations, can log their feedback on the quality of search results.

Full details of current and future plans for retrieval evaluation are outlined in the report on Milestone 7.4.

T7.3 Translation Evaluation

As the translation quality is the key output of the project, a meticulous framework has been implemented by which the performance of the MT systems can be assessed.

Using the evaluation environment of Cross Language, translation output is evaluated by human assessors on the following criteria:

- *Adequacy* – translated segments given a score from 1—5;
- *Benchmarking* – output from multiple systems is ranked. This allows us not only to compare our MT system with others, e.g. Google, but it also allows us compare different versions of our own systems, e.g. general MT vs. domain-specfic MT vs. MT/TM integrated system;
- *Error categorisation* – classifying errors in the output into different categories, e.g. grammar, style, semantics, etc.
- *Post-editing effort* – time taken to correct MT output

Additionally, translation output is evaluated using automatic measures such as BLEU and METEOR. This evaluation is more developer-centric and allows us to compare different iterations of MT systems in a more effective manner.

All of the above evaluations are carried out on predefined representative test sets. For each translation direction, there exists 8 test sets and reference sets according to the patent domains in the IPC classification. This breakdown of the evaluation process allows us to pin point those types of patents with which the MT engines perform particularly well/poorly.

While some preliminary results can be found in Deliverable 5.1, a full report on the first evaluation cycle will be submitted in M18 (October 2011) as a requirement of Deliverable 7.6

2.6.4 Use of Resources

Beneficiary	PMs Actual (Year 1)	PMs Planned (3 years)
DCU		
ESTeam		
IRF		
Cross Language		
WON		
Total		

2.6.5 Summary

The main developments relating to evaluation in the project have been highlighted in the section. Outside of the scope of the Description of Work, the En—Pt was evaluated by the EPO and the Portuguese National Patent Office.

Additionally, frameworks for the evaluation of search and translation quality have been designed with a large portion of analyses already carried out in preparation for deliverables in October 2011. This includes the delivery of the evaluation survey as part of Deliverable 7.1 which was been prepared in conjunction with the WON user group.

2.7 WP8 Dissemination

2.7.1 Objectives

The Dissemination work package is concerned with the communication of all PLuTO related materials – news, scientific findings, events, activities – in order to achieve and maintain high visibility for the project. This will afford PLuTO greater opportunities to engage with interested communities and users groups, foster collaborations, and ultimately increase the overall impact and long-term sustainability of the software platform. Deliverables and milestones falling due in the period are shown in Table 7.

Mi8.1	Dissemination and training plan and website available	☑
D8.1	Dissemination plan	☑

Table 7 Milestones and deliverables due between M1 and M12

2.7.2 Progress Highlights

A significant ramp-up effort was required for dissemination in the early stages of the project which sets the stage for coming years. In Month 6, a dissemination plan and budget was delivered which sets out the key events which the consortium is targeting over the duration of the project it terms of attending, presenting/exhibiting prototypes, sponsorship and technical publications.

Related to this, PLuTO has already been presented at a number of conferences across Europe and North America and has had two peer reviewed conference papers accepted. These are listed below in section 4.5

Finally, the project website – http://www.pluto-patenttranslation.eu – was launched in July 2010. Along with the PLuTO Facebook page and Twitter tag #PLUTO_EU, it serves as the public face of the project and is constantly updated with news items and relevant downloadable content.

2.7.3 Tasks

T8.1 Training

At the project kick-off meeting and subsequent technical meetings, project partners were introduced to one another's software through demonstrations and discussions. This will continue and in the future, where appropriate, project members will travel to the sites of other partner to work on specific problems. In terms of external training, to date there has been no actual software on which to train users. However, from the very beginning of the project's second year following the meeting with the WON working group, there have been initial steps taken to introduce our first system prototype to users. This will continue at events over the course of 2011.

T8.2 Dissemination

This task relates to the development of a strategy within the project for promoting key developments and results to interested communities. The traditional channels for this are via the project website and through event organisation and attendance. As mentioned, the project website has been live since the early part of year one, while PLuTO has been widely exposed at a number of events. These items are described in greater detail in sections 4.4 and 4.5 respectively.

In addition to events which have already take place, Deliverable 8.1 Dissemination Plan from month 6 outlines the events which PLuTO will be targeting over the course of the entire project. These include a mix of MT and IP conferences, both academic and business in nature, where participation is envisaged in a number of forms such as peer reviewed publications, demonstrations/exhibitions, invited talks and sponsorship activities. These events have been budgeted within the scope of the project resources and will allow us to maximise the visibility of the project output.

2.7.4 Use of Resources

Beneficiary	PMs Actual (Year 1)	PMs Planned (3 years)
DCU		
ESTeam		
IRF		
Cross Language		
WON		
Total		

2.7.5 Summary

Ramp-up activities in the context of dissemination were successful with the project web site being deployed as well as a number of introductory presentations and press releases. This has continued with additional publications and presentations as well as a concrete dissemination strategy to see the project through to its conclusion.

2.8 WP9 Exploitation and Standardisation

2.8.1 Objectives

The work package on Exploitation and Standardisation of PLuTO is charged with keeping the consortium in touch with current market trends, in terms of both technical and commercial developments, in the area of translation service provision tools, particularly as relates to patents. Additional, a strategy will be developed to exploit the results of the project via the most appropriate distribution channels. Deliverables and milestones falling due in the period are shown in Table 8.

Mi9.1	Report on joint SME feasibility available	☑

Table 8 Milestones and deliverables due between M1 and M12

2.8.2 Progress Highlights

The main highlight to date has been the delivery of a feasibility report on the formation of a joint SME based on the outputs of the project. This includes an initial market observation exercise of the patent/IP landscape, as well as potential locations, user scenarios, and risks of a venture in this area. Important information has also been gained through discussions with WON, our advisory board, and project-external patent experts.

Significant insight too has been gained through the consortium's collaboration with the EPO, particularly as relates to the direction in which that body is going in the future as relates to their patent translation requirements.

2.8.3 Tasks

T9.1 Market Observation

First investigations suggest multiple feasible setups for commercialisation and exploitation based on the outcomes of the project. While a joint SME would entail a combination of data resources, search technologies and automatic translations, the combination can also be split into two very profitable separate spin-off companies for the data resources and search technologies and another one catering for automatic translations.

While the patent search market is quite heavily saturated, the collaboration with the EPO to date has not only served to indicate the need for patent *translation* technology but also highlights the commercial viability and value of the PLuTO software. It has also provided us with invaluable market research data from their end included those language pairs most in demand, e.g. Japanese and Portuguese. We can leverage on this information to direct and readjust our work plan where necessary.

More recently, the initiatives of Google in the patent translation market have created questions as to the viability of being able to compete as an SME providing specialist translation software, particularly as Google intend to provide these services for free (or at least in exchange for data). These fears have been eased somewhat by the fact that Google provides a non-secure environment that, according to our investigations, patent lawyers as users do not trust. However, developments at Google in this area are certainly something to keep a close eye on.

In addition to Google, other patent translation service providers are continuously monitored in terms of the scope and quality of the services they provide. These include the EPO's Espacenet service, the WIPO's Patentscope service, and the translations services offered by the various national patent offices around the world, e.g. the K2E-PAT service of the Korean patent office.

Full details on market observation activities as well as a feasibility study into the formation of an SME can be found in the report on Milestone 9.1.

2.8.4 Use of Resources

Beneficiary	PMs Actual (Year 1)	PMs Planned (3 years)
DCU		
ESTeam		
IRF		
Cross Language		
WON		
Total		

2.8.5 Summary

Initial market observation has suggested that pressing ahead with planning for a joint SME based on the output of the project is a feasible option, particularly for translation. Our collaboration with the EPO has validated our technical approach to MT as well and the need for such a service. However, the consortium must be cautious of patent translation initiatives at Google at adapt the work to remain competitive if necessary.

3 Deliverables and Milestones Tables

Table 1. Deliverables

Del. no.	Deliverable name	WP no.	Lead participant	Nature	Dissemination level	Due delivery date from Annex I	Delivered Yes/No	Actual / Forecast delivery date	Comments
1.1a	Annual Project report	1	DCU	R	P	31/03/11	Yes	14/04/11	Due 60 days after the end of the reporting period. Delivered in advance of the site review.
2.1	Data corpora and data standards v1	2	IRF	O	RE	31/03/11	Yes	14/04/11	
3.1	First web application and user interface	3	EST	O	CO	31/03/11	Yes	14/04/11	
4.1	Translation memory components for 2 languages	4	EST	O	CO	31/03/11	Yes	14/04/11	
5.1	MT engines for 2 language pairs	5	DCU	O	CO	31/03/11	Yes	14/04/11	
6.1	First prototype	6	EST	D	RE	31/03/11	Yes	14/04/11	
7.1	Survey Structure	7	CL	O	P	31/03/11	Yes	14/04/11	
8.1	Dissemination Plan	8	IRF	R	CO	31/09/10	Yes	25/10/10	Delayed with approval due to upcoming project meeting

Table 2. Milestones

Milestone no.	Milestone name	Due achievement date from Annex I	Achieved Yes/No	Actual / Forecast achievement date	Comments
2.1	(EN-DE) patent data available	30/06/10	Yes	30/06/10	DE was replaced by PT
2.7	Metadata annotation sets and guidelines available	31/09/10	Yes	25/10/10	Integrated with Mi2.7
4.1	TM Resources for EN-DE	31/09/10	Yes	25/10/10	DE was replaced by PT
5.1	MT Engine for EN-DE	31/09/10	Yes	31/09/10	DE was replaced by PT
6.1	Integration Contracts complete	31/09/10	Yes	25/10/10	
8.1	Dissemination and training plan and website available	31/09/10	Yes	25/10/10	
2.2	FR patent data available	31/12/10	Yes	06/07/10	
2.8	Final metadata annotation sets and baseline selection engine	31/03/11	Yes	14/04/11	
4.2	TM resources for EN-FR	31/03/11	Yes	14/04/11	
5.2	MT Engine for EN-FR	31/03/11	Yes	14/04/11	
6.2	Integration Prototype Available	31/03/11	Yes	14/04/11	
7.1	Survery structure and content v1	31/03/11	Yes	14/04/11	
7.4	Integrated cross-lingual evaluation framework available	31/03/11	Yes	14/04/11	
9.1	Report on joint SME feasibility	31/03/11	Yes	14/04/11	

4 Project Management

4.1 Management Tasks and Achievements

In this section, we describe work carried out over the course of the period of which the project coordinators were directly responsible. Additionally, we present tasks performed and other achievements that were not specifically outlined in the Description of Work.

4.1.1 EPO Collaboration

Timeline

Following a PLuTO project briefing to the European Patent Office (EPO) Machine Translation project team at Den Haag in June, a request was received by the PLuTO project co-ordinator 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.

The MT service went live through the EPO's Espacenet service in September 2010 and was in use until March 2011. During this time, almost 15 million words were translated at an average speed of approximately 2,800 words per minute. A summary of the resources expended by the service over the course of the collaboration is shown in Table 9

What	En—Pt	Pt—En
Documents translated	11,127	4,399
Words translated	13,915,782	690,592
Average speed (words per second)	51.7	41.5
CPU time spent (seconds)	943,888	92,537

Table 9 Statistics from the EPO translation service

The service was suspended by PLuTO in March 2011 for a number of reasons.

An initial condition of the collaboration was that PLuTO translation quality must meet certain criteria before being deemed acceptable for use in Espacenet. In summary, the criteria were to achieve 3.0 or greater (out of 5.0) on a human evaluation scale and to be equal or better than Google translation quality. The service was evaluated at the Portuguese national patent office before deployment and was deemed to be of sufficient quality to use by the EPO. The PLuTO system scored **3.5** for Portuguese—English and **2.9** for English—Portuguese. This judgement was made on the proviso that PLuTO would continue efforts to improve the English—Portuguese MT engines with the help of the EPO in terms of data provision.

We continued development of the MT systems with an agreed date of January 30 2011 to provide improved systems which was met. However, at this stage, unbeknownst to us, the EPO had 're-evaluated' the original systems using the same evaluation framework but with different reviewers; this time in-house EPO employees were used to judge translation quality. Following this evaluation, both PLuTO systems were scored **2.8** and deemed unfit for production. We questioned the EPO on the integrity of this evaluation as we considered it to be very inconsistent not only with their previous evaluation, but also with standard practices for this type of task. However, no satisfactory response was received.

This unexpected evaluation also coincided with the EPO's signing of a memorandum of understanding with Google for the provision of MT services and a conspicuous reduction in communication on the part of the EPO with the PLuTO coordinator.

Furthermore, the project was originally undertaken with an agreement in place that the EPO would ultimately provide some payment for the services provided. This was based on a proposed tender for MT services to be released by the EPO in late 2010 which ultimately never materialised. While the EPO initially looked into other means for funding this project, it was clear that nothing would be forthcoming as soon as they entered into dealings with Google.

In the best interests of the project, we decided to discontinue development on the EPO MT service and to refocus our efforts on project specific tasks, using the experience with the EPO to help drive the development of our technical solutions and business models with a heavy user-driven focus.

Benefits

Nevertheless, the collaboration has provided a number of distinct technical benefits for the consortium. Firstly, the consortium received a large quantity of patent data, some of it of a very high quality, which satisfied some requirements of work packages 2 and 4. In building MT engines for the EPO and developing a full-scale production-level web service through which the translation service was delivered, we satisfied some requirements of work packages 3 and 5. When preparing hosting arrangements in DCU for the web service, we developed a number of APIs and interfaces which now serve as core elements of the integration contracts detailed in work package 6. Furthermore, upon implementation of the translation service, it was initially thoroughly evaluated 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. This provides the partners with a substantial head start when it comes to addressing some of the evaluation questions in work package 7.

Finally, the consortium has been able to capitalise on the collaboration as a means to further promote the project and to satisfy some of the requirements of work packages 8 and 9. Firstly, the technical aspects of the work form the core of the publication accepted at the AMTA conference (sees section 4.5). Additionally, the service the consortium provided to the EPO has been publicised on both the website of the project as well and the EPO and was referenced in a number of high profile presentations throughout Europe.

4.1.2 Quality Management

The coordination team instituted a quality management system within the project to ensure on time submission of all milestones and deliverables, as well as consistency of output in terms of quality, content and style.

Under this system, each document must pass a 3 point quality check prior to submission. One month before the due date, a draft of the deliverable is reviewed by an assessor external to the work package from which the deliverable originates. Following that, feedback is taken into account and the document is finalised by the work package leader. The last step involves the project coordinator signing off on the deliverable before submission.

This process was first employed for the deliverables submitted for M12 and was relatively successful. Some diversions took place, typically due to the (un)availability of individuals at certain times, but this ultimately did not affect any deliverables. Further details on this process can be found in the Deliverable 1.0a Quality Management Policy submitted as an M6 deliverable.

4.1.3 Reporting and Communication

Reporting

In order to ensure project activities are being carried out in a timely manner and to the high standards expected of the project, an internal periodic reporting procedure has been implemented. Each partner in the consortium provides a short monthly progress report to the project co-ordinator documenting the progress made in the preceding month towards the objectives of each work package in which the partner in question is involved.

Additionally, a quarterly report is submitted per work package to the project co-ordinator focussing on progress made with respect to upcoming milestones and deliverables. Partners also submit a quarterly financial report in order to assist the coordinators in compiling the annual financial report.

While the quarterly reports have been very beneficial, we noticed that often for the monthly reports there is little significant difference from month to month to justify such a frequent report. Given this, we plan to replace this report with a monthly Skype call with a representative from each partner and a bimonthly call between the coordinator and the technical coordinator.

Full details on the original reporting procedures can also be found in the aforementioned Deliverable 1.0a.

Communication

Communication between the various project partners has been quite active throughout the course of the first year. Individuals have used a number of different channels where appropriate to hold meetings, share documents and data, and to generally keep in touch.

The project website (described fully in section 4.4) was implemented with a member's section accessible only by project partners. This acts a kind of wiki whereby draft documents can be shared and edited amongst partners. It also serves as an archive for deliverables, meeting minutes, presentation slides, and other such information which should always be available to all partners yet remain out of the public domain.

The project mailing list (pluto@ir-facility.org) has also been used extensively by all members and a number of smaller project meetings have been held via Skype and using teleconferencing facilities. Additionally, partners have taken advantage of occasions where they have been attending external events individually to set some time aside to discuss project related matters.

4.1.4 Risk Management

A risk management plan has been developed to identify potential threats to the project and set out measures to reduce their impact and manage the fall-out of an identified risk

scenario should it occur. Risks are assessed under a number of headings including likelihood of occurrence, impact should it occur, contingency plans etc, and are ranked according to their overall severity. The full risk management document was submitted in M6 as Deliverable 1.0b.

4.1.5 Advisory Board

At an early stage in the project, 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. PIUG. To that end, the following individuals accepted invitations to our advisory board by the end of 2010:

Dr. Fred Hollowood

Fred Hollowood, the director of MT research and development at Symantec Corporation, will act as the MT expert. Fred is also associated to DCU/CNGL through Symantec's involvement as an industrial partner and has widespread experience in the exploitation of MT in an industrial/commercial setting.

Mr. Viggo Hansen,

Viggo, a management consultant, will serve as our IP expert. Viggo is a particularly appropriate choice as he also has a translation background and was a pioneer when it comes to exploitation of MT for patent translation having founded Lingtech and served as CEO of Hofman-Bang A/S and Zacco A/S. Viggo is also a former president of the European Association for Machine Translation and is the founder and organiser of the IPWare Summit conference.

Dr. Greg Grefenstette

Greg, the chief scientific officer at Exalead, will act as our IR expert. Prior to taking his currently position at Exalead, the world's largest commercial search engine outside of Google, Yahoo, and Microsoft, Greg was an active researcher in the field of IR having worked at Xerox Grenoble and in the USA.

Mr. Stephen Adams

Stephen, the managing director at Magister Ltd., is also one of our IP experts as well as representing a user base. Prior to founding Magister, an IP consultancy, Stephen worked as a technical patent searcher in the area of chemistry. Between 2002 and 2006, Stephen was also the director-at-large of the PIUG and received a Special Recognition Award for this in 2008.

The consortium met with the advisory board for the first time at our AGM in Amsterdam on April 5th 2011. A very fruitful day was spent discussing the progress of the project to date as well as plans going forward. The day also included a dry-run of the review process whereby individual work packages and deliverables were discussed in detail.

4.2 Project Meetings

4.2.1 Kick off meeting – Vienna – 05-06/05/10.

The project kick-off meeting was held at the IRF offices in Vienna. In addition to PLuTO Project consortium members, this meeting was attended by the PLuTO project officer, Susan Fraser, and representatives from the EPO. Due to the grounding of flights from Ireland

because of the Icelandic ash Cloud, representatives from DCU attended the meeting via video conference.

4.2.2 Technical Meeting – Den Haag – 21-22/06/10.

Member of the technical teams from all core partners met in Den Haag on June 22nd. The goal of this meeting was to discuss initial technical details across partners and put in place a plan for the completion of all first year deliverables.

This meeting 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.

4.2.3 General Assembly – Dublin – 18-19/10/10.

A full two-day general assembly was held at the CNGL offices in Dublin City University to review progress in all work packages over the course of the first six months of the project. Month 6 project deliverables were also reviewed and discussed prior to their submission to the EC. Finally, a more in-depth technical working session was held to agree integration contracts between components of the PLuTO system and create a plan for the delivery of the first prototype at month 12.

4.2.4 AGM – Amsterdam – 04-05/04/11.

The first PLuTO annual general meeting was held in Amsterdam shortly after the official end of year one date and all five partners were represented. The location was chosen in order to collocate with WON's AGM. This allowed members of the consortium working on WP7 Evaluation to meeting the WON's PLuTO working group to established plans for year 2 when WON's involvement in the project increases.

At the AGM itself, all M12 deliverables as well as the annual report were reviewed and well as the establishment of tentative plans for year 2. This meeting also marked the first time the advisory board met with the consortium to provide their direct feedback on progress and act as assessors for a dry run of the year one review process.

4.2.5 WON Meeting – Weesp – 06/04/11

Representatives from the technical partners, specifically those involved in evaluation, met with members of WON following the AGM. This meeting was primarily focussed on the preparation of the evaluation survey for Deliverable 7.1 and also to establish the various points in the workflows of patent users where translation is required to facilitate planning in work package 9.

4.3 Deviations from Plans

Following the collaboration with the EPO, and consultation with the PLuTO project officer, the languages treated in this project period were adapted as shown in Table 10. This deviation from the original work plan has been driven by the opportunity to take advantage of an external customer's requirements in a pilot project. At the project kick-off meeting, this area of the project (the languages treated) was identified as one where the consortium could be relatively flexible, particularly as related to reacting to market demands and customer requirements. Over the course of the remainder of the project, we envisage that the languages may change one more, particular given the noises we are hearing from patent

users regarding the need for a high quality MT solution in Asia, particularly for Japanese and Chinese.

Year 1 Prototype *Original* Languages	Year 1 Prototype *New* Languages
French	Portuguese
German	French

Table 10 Changes in language pairs for year 1

4.4 Website Development

A placeholder website containing the information found on the PLuTO fact sheet was live at http://www.pluto-patenttranslation.eu shortly after the kick-off meeting. The full website in its current format went live on 15/07/10. It is hosted by the IRF and jointly administered by the IRF and DCU. As of 18/04/11, the website has received 1,817 unique hits – an average of around 50 visits per week.

The website contains overviews of the project, its goals and the consortium. Information on specific work packages is also presented. There is a download section where users can access selected documents, presentations and other information that is for public consumption. The is a latest news bar which provides up to date information on project related activities and events which PLuTO partners will be attending.

Finally, there is a member only section of the website for partners to log in. This acts as the project wiki whereby we can share internal reports, deliverables, minutes, templates and other data that is not for public consumption.

In future, the project website will also serve as an entry point to the PLuTO integrated prototype.

4.5 Dissemination Activities

PLuTO members have attended a number of events over the course of the first project period in which the project has been presented in a number of formats. These are summarised below:

May 2010 **EAMT St-Raphael, France**	The conference of the European Association for Machine Translation, EAMT 2010, held a special FP7 showcase plenary session to highlight those projects exploiting MT technologies. PLuTO was introduced during a short oral presentation and then during a poster presentation. There was much interaction with attendees and, in fact, one of the interested parties, Viggo Hansen, the local host of the conference, now serves on the PLuTO advisory board.
June 2010 **IRFS Vienna**	PLuTO was presented in the Exhibition hall of the IRF Symposium. A description text and logo were placed on a backlit column of about 3m height where the IRF presented the project. PLuTO was also mentioned in the programme of the symposium and of the IRF Scientific Conference
July 2010 **SIGIR**	PLuTO was exhibited as a poster at the IRF's booth in the exhibition hall ACM SIGIR. Keen interest was shown by participants at this event, the most important IR conference in the field, particularly as relates to the

Geneva	commercialisation plans for the project.
October 2010 **AMTA** **Denver, CO**	A PLuTO paper entitled "*PLuTO: MT for Online Patent Translation*" was accepted for publication at the conference of the American Machine Translation Association, AMTA, in Denver. The paper was presented during the user track of the conference by members from DCU and was well received.
October 2010 **CKIM** **Toronto**	A PLuTO poster was displayed during sessions at both the Patent IR workshop at the CKIM conference in Toronto, and the main conference itself. Representatives from the US patent office engaged project partners in interesting discussions.
October 2010 **ICIC** **Vienna**	The PLuTO project was presented at the ICIC Conference in Vienna. Te IRF had bought a space in the exhibition hall and distributed PLuTO fact heets along with the IRF communication materials. As an exhibitor, the IRF had a short presentation in the plenary meeting where the PLuTO project was introduced to the audience.
November 2010 **Microsoft** **Dublin**	A PLuTO poster was exhibited at the Localisation Innovation Showcase in Dublin. This event was hosted by Microsoft Ireland, an industrial partner of the CNGL in DCU, and attendees were a mix of academics from within the CNGL as well as invited guests of Microsoft from relevant industry.
April 2011 **WON AGM** **Utrecht**	PLuTO representatives from Cross Language and the IRF attended the WON AGM, presenting PLuTO as a guest talking during the single day event. This followed a day of fruitful discussions with WON where the evaluation survey of Deliverable 7.1 was created. The contents of this survey were presented to members at the AGM for their feedback and suggestions.
April 2011 **EAMT** **Leuven**	A PLuTO paper entitled *"Experiments on Domain Adaptation for Patent Machine Translation in the PLuTO project"* was accepted for oral presentation at the user track of EAMT to take place in May 2011. PLuTO will also be presented as a poster during special session on EC funded projects

In addition to conferences and workshops, a PLuTO fact sheet was also released towards the beginning of the project along with press releases in both English and German. A PLuTO Facebook page was also set up and the hash tag #PLUTO_EU is being used on Twitter to highlight PLuTO related posts. Both of these channels are advertised on the project website and users have the option to 'share' news items via a toolbar at the top of each story.

Finally, PLuTO members have attended a number of other conferences and events over the course of the year in their non-PLuTO roles and have used these opportunities to spread information on the project.

4.6 Issues Arising

4.6.1 EPO Collaboration – Time Management

There were issues in terms of the time that the EPO related work was taking away from DCU's coordination activities. This was rectified through discussions with the EPO and the EC in order to allow us to prioritise project specific work and the EPO collaboration more effectively.

4.6.2 IRF Departure from Consortium

Following the year one review, the consortium will be reduced to four partners when the IRF withdraws from the project due to their financial situation. We intend to take the positives from this affair and use it as an opportunity not only to redistribute the tasks and resources of the IRF across the remaining partners, but also to refocus the work based on the feedback we have received from our advisory board and users over the first period.

This will involve reducing the role of search in the project and putting increased efforts and resources into machine translation and developing specific features and applications. We envisage an increased technical role for Cross Language in this reallocation not only because the consortium would be quite top heavy with the majority of tasks and resources falling on DCU and ESTeam, but also because of their significant technical capabilities.

We will take these decisions following advice and feedback from our expert reviewers and the project officer at our annual review.

A number of other considerations are already being addressed, including relocation of the search servers, the project website, mailing list, and physical acquisition of the data.

4.6.3 Andy Way Leave of Absence

Prof. Andy Way will be taking a three year leave of absence from DCU which will in effect see out his involvement in the PLuTO project. A replacement for his position within DCU is already actively being sought and the application process closes on May 6th 2011. It is hope that the replacement professor will be in place by the time Andy departs. However, even if this is not the case, we do not believe this will affect the work in the project in any significant manner as the expertise on the ground in DCU exists to compensate sufficiently for Andy's absence.

4.6.4 Hiring

As coordinator, DCU was the only partner in the consortium that required new full time hires to being work immediately on the PLuTO project. To that end, there were some delays in the ramp up on staffing. While Dr. John Tinsley was hired as full time-coordinator from the beginning of the project, and Dr. Declan Groves worked on technical development, a number of positions remained open. Two of these positions were filled in time for the first general assembly in Dublin in October through the hiring of Alex Ceausu and Jian Zhang, while others remain open. It is planned to leave these positions open for the time being for a number of reasons particularly pertaining to the budgetary restrictions of the PSP funding model.

5 Future Plans

Core deliverables

While there will be some level of adjustment to future plans given the reallocation of the IRF's workload, the principal objectives of the project remain in place. Through ongoing provision of data in year two, machine translation engines and translation memory resources will be developed for German and Dutch languages, and these will be incorporated into the online prototype. We remain flexible on the possibility of changing or adding new languages, particularly Asian languages, depending on the feedback from users in our evaluation survey.

Regarding the survey, its findings will serve as a significant springboard for a number of activities, particularly as relates to the preparation of an Exploitation Plan as Deliverable 9.1 in M18. Furthermore, the findings of the survey will help define a number of features to be integrated into our prototype and translation services which will add significant value to our services.

Aside from the findings of the survey, deliverables will be submitted on the quality of MT and search.

Reallocation of IRF resources

A point of immediate importance following the year one review is an official submission to the EC on the reallocation of responsibilities and resources from the IRF to the remaining partners. As mentioned in section 4.6.2, our aspiration based on feedback from our advisory board and users is to refocus our work somewhat from search and translation to a wider-coverage translation service with valued-adding user-requested features. However, we intend to first consider feedback and suggestions from our expert reviewers and the EC, and consult with them, before we take our final decision.

Dissemination activities

PLuTO members will continue to be proactive in their dissemination activities over the course of the second year, particularly as relates to the requirements of Deliverable 8.2. A number of plans are already in place, including:

- The aforementioned PLuTO paper and poster at **EAMT** in Leuven, May 2011;
- PLuTO has been accepted for an oral presentation during a session at the **PIUG (Patent Information User Group) conference** in Ohio, May 2011;
- PLuTO is an exhibitor at the reduced version of the **IRF Symposium** in Vienna, June 2011;
- PLuTO has been invited to speak at the **WIPO Symposium** in Geneva, September 2011;
- Finally, several PLuTO members intend to attend the **IPWare Summit** in San Remo, October 2011, where an invitation stands for an oral presentation as well as the possibility to exhibit our prototype system.

6 Summary

[to be completed]

7 Bibliography

Ceausu, A., J. Tinsley, J. Zhang, A. Way, and P. Sheridan. 2011. Experiments on Domain Adaptation for Patent Machine Translation in the PLuTO project. In *Proceedings of the 15th Annual Conference of the European Association for Machine Translation*. Leuven, Belgium.

Tinsley, J., A. Way, and P. Sheridan. 2010. PLuTO: MT for Online Patent Transaltion. In *Proceedings of the 9th Conferences of the Association for Machine Translation in the Americas*. Denver, CO.