Project no. 215231

TrebleCLEF

Treble-CLEF: Evaluation, Best Practices and Collaboration for Multilingual Information Access
IST: ICT-1-4-1, Digital libraries and technology-enhanced learning

Deliverable 6.5
Exploitation Plan

Start Date of Project: 01 January 2008
Duration: 24 Months

Organisation Name of Lead Contractor for this Deliverable: CELCT

Version: Final

Project co-funded by the European Commission within the Seventh Framework programme

---

Document Information

Deliverable number:	6.5
Deliverable title:	Exploitation Plan
Due date of deliverable:	12|09
Actual date of deliverable:	December 2009
Author(s):	Danilo Giampiccolo, Nicolas Moreau, Nicola Ferro
Participant(s):	ALL
Workpackage:	6
Workpackage title:	Dissemination
Workpackage leader:	CELCT
Dissemination Level:	PU (Public)
Version:	[1]
Keywords:	exploitation, dissemination, portal, evaluation package,

History of Versions

Version	Date	Status	Author (Partner)	Description/Approval Level
0	20/11/09	Outline	Danilo Giampiccolo	Distributed for comments
1	18/12	First draft	Giampiccolo, Moreau	Distributed for comments
2	21/12	Near-final	Giampiccolo, Moreau, Ferro	Distributed for approval by Consortium
3	22/12	Final	Giampiccolo, Moreau, Ferro	Approved

Abstract

This document describes the exploitation plan of the TrebleCLEF Coordination Action which aims at proposing solutions for sustainable follow-up activities in the interested research and application communities after the project's lifetime. After giving an overview of the outcomes of the activity carried out within the framework of the project, the initiatives aimed at exploiting the knowledge and resources produced are described in detail.

Table of Contents

Executive Summary

The aim of the TrebleCLEF Coordination Action has been to support the development and consolidation of expertise in the multidisciplinary research area of multilingual information access (MLIA) and disseminate this knowhow to the application communities through a set of complementary activities. This documents lays out the plans for post-project exploitation of the main results of TrebleCLEF. It first provides a brief overview of the principal outcomes of the project, including a brief summary of the dissemination activities. The ways in which we intend to further exploit these results after the end of the project lifetime are then described.

The knowledge and expertise acquired through the activities and the research carried out during TrebleCLEF have been disseminated through a considerable scientific production, consisting of articles and papers that the members of the consortium have published in journals, proceedings and specialized publications. The Proceedings of the two campaigns of the Cross Language Evaluation Forum (CLEF2008 & CLEF2009), organized within the framework of TrebleCLEF and prepared for publication in the Springer LNCS series, have significantly contributed to making the results of experimentation in MLIA application development and evaluation available to the public.

The various workshops, tutorials and the Summer School organized by the project have also contributed to knowledge dissemination in various MLIA-related sectors. A particularly important role has been played by the three Best Practice reports which provide recommendations for system designers and developers in the Multilingual Information Access (MLIA) domain with respect to tools and languages resources, system implementation, user requirements and system evaluation. These guidelines have been presented in a variety of forms, both as tutorials and oral presentations. They were also presented at a dedicated event, the MLIA Technology Transfer Day, organized with the aim of bringing together members of the R&D and applications communities to exchange information and discuss, on the one hand, the latest research advances and, on the other, the actual requirements of developers and users.
Throughout the project, the TrebleCLEF website has been incisive in making publicly available information on project events and initiatives, and publishing documents related to project activities.

In the Exploitation Plan, we focus on three major outcomes produced by the TrebleCLEF Consortium: the Best Practice Portal; the DIRECT scientific digital library; CLEF evaluation packages.

Best Practice Portal

The first initiative consists of the Best Practices Portal, which has been set up to present the knowledge presented in the Best Practices reports in an easily accessible and systematically organized online repository. The Portal was made publicly available at the end of November, via the TrebleCLEF website. Its main objective is to provide a valuable source of information and tools for application development and evaluation and this is reflected in the architecture, which presents four top-level categories corresponding to the Best Practices recommendations. However, the Portal is a valid means to share other MLIA related data and resources. Users are encouraged to actively contribute to its population by adding new content after registering. The Portal -designed by CELCT, which also will be in charge of its maintenance and of the supervision of the content after the end of the project - has been advertised on several IR and NLP mailing lists, and will be further promoted through periodic reports on newly added content. The CLEF 2010 conference, to be held in Padua September 2010, will represent an important occasion both to test the Portal and to stimulate participants in the campaign to collaborate to the portal population, sharing their own resources within the MLIA R&D community.

DIRECT (Distributed Information Retrieval Evaluation Campaign Tool)

The second exploitation activity concerns the DIRECT system[1], which has been designed and developed by UNIPD. DIRECT is a digital library system which manages and makes accessible the scientific data produced during an evaluation campaign, e.g. runs, performance measures, pools, statistical analyses, and so on.

DIRECT has been adopted to manage some of the core CLEF tracks (Ad-hoc, Domain-specific, GeoCLEF, Grid@CLEF, ImageCLEFPhoto) since 2005 and provides access to: more than 5.6 million documents; more than 1.5 million relevance assessments for more than 600 topics made by over 200 assessors in 15 different countries; more than 2,500 experiments, which amount to about 117 million tuples, submitted by over 170 participants from about 30 different countries; over 5.5 million performance measures and descriptive statistics; and, about 20,000 plots and statistical analyses graphs.

Upon agreement with UNIPD, DIRECT will be made available after the end of the project to host and manage to host and manage tracks in future evaluation campaigns that request such support. DIRECT will also be used to make data and information accessible online, upon registration; this will foster the re-use and exploitation of the experimental results and ease the comparison with the experimental evidence gathered during the different CLEF campaigns. Note that the document collections will be used internally in DIRECT but will not be available for online download, since they are regulated by the copyright agreements with data providers managed by ELDA. DIRECT could provide pointers to the ELDA catalog for this purpose.

CLEF Evaluation Packages

The third exploitation activity regards the CLEF evaluation packages created and distributed by ELDA. ELDA has produced evaluation packages containing a variety of reusable resources (document collections, test queries, relevance judgements, tools, etc) collected during several CLEF evaluation campaigns. As these evaluation packages are very valuable outputs of the project, ELDA has prepared an exploitation plan which aims at maximizing the possibility of access to such resources by users both from the IR and NLP communities. The goal is to make the CLEF resources publicly available as a set of task evaluation packages, covering all the CLEF evaluation campaigns, from 2000 to 2008. All the evaluation packages which have been completed will be included in the ELDA catalogue by January 2010, and more will be added during the first quarter of 2010, provided the successful negotiation of the use of data. It is hoped that it will be possible to add data from the CLEF 2009 campaign to the existing packages after the end of the project.

As far as distribution is concerned, ELDA is working on securing a fair distribution procedure on behalf of the CLEF consortium, namely establishing a simplified licensing scheme and exploiting a regular distribution channel, such as the existing online ELDA/ELRA resource catalogue.

Moreover, in order to actively disseminate the evaluation packages ELRA/ELDA will be make full use of its dissemination channels (ELRA newsletter, LREC conference, mailing lists, etc.) to promote awareness of the availability of the reusable language resources and the evaluation procedures and methodologies created by CLEF.

[1] http://direct.dei.unipd.it/about.html

1 Introduction

The primary goal of the Treble CLEF project has been to disseminate knowledge and spread excellence in the domain of Multilingual Information Access (MLIA). Thus a variety of initiatives for knowledge dissemination in research and industry communities has been proposed throughout the duration of the project in order to exploit the results achieved in the research activities and the tools and data collected and produced by TrebleCLEF. In this Introduction we briefly list the main results.

The Best Practices recommendations for MLIA language resources (Deliverable D5.2), System oriented and User-Oriented MLIA (Deliverable D 3.3), and Test collection creation, evaluation methodologies and language processing technologies (Deliverable D4.2) represent one of the most important outcomes of the project and have been already disseminated through a number of initiatives aimed at presenting the Best Practices to the public, such as workshops and tutorials (see the Progress Report of D1.1.2 for details).

The Best Practices recommendations also played an important role in the MLIA Technology Day, held in Berlin, December 2009, which aimed at bringing the players in MLIA domain (researcher, developers and end users) together to discuss and exchange ideas on how research and industry can in interact and how technologies developed by researchers can be used in real applications in MLIA. The good response to the event confirms the interest of both research community and industry to share common knowledge and expertise.

Beside the Best Practices recommendations, a considerable volume of scientific production relevant to the TrebleCLEF project has been also produced and divulgated in the community both through scientific publications and presentations at conferences and workshops. The list of the publications and of the events attended by members of the TrebleCLEF consortium is given in the Progress Report, and is also published on the project's website.

The two CLEF evaluation campaigns supported by TrebleCLEF have also represented an important source of MLIA related data, tools and resources. The main means for dissemination of the results achieved both in the development of applications and in the creation of tools and methodologies for their evaluation have been the Post-Workshop proceedings published by Springer. The Proceedings of the CLEF 2008 campaign were published in September 2009, meanwhile the CLEF 2009 Proceedings are being editing and will be published in September 2010.

Another major resource to disseminate the project's outcomes has been the TrebleCLEF website (www.trebleclef.eu), which has played an important role in making the documentation and information relevant to the project available for a wider public, by providing information about the project, advertising events and initiatives, and publishing the list of scientific publications. A password protected area is available for consortium-confidential documents. The website will be kept online at least for one year after the end of the project to ensure that all the valuable information stored will remain available for the public.

Of the vast amount of knowledge, know-how tools and resources mentioned about , some outcomes appear to be particularly suitable to be further exploited after the end of the project, namely the Best Practices recommendations; the evaluation methodologies and the results of the data analysis produced in the CLEF campaigns; and the CLEF evaluation packages.

To maximize the circulation and exploitation of these resources among the MLIA community, the following initiatives are planned:

- making available the content of the Best Practices through a dedicated portal. The Best Practices Portal, whose prototype was already presented at the 12 month Review meeting, was officially released at the end of November 2009 for public access. Its main goal is to make the content of the Best Practices produced during the project publicly available in a easy and dynamic way. The Portal has been in fact designed to be updated also by the contribution of users, who can add new content after registration. This means that besides the information taken from the Best Practices a information on variety of MLIA related data, tools, and resources can be shared through the Portal, making it a useful reference instrument not only for MLIA researchers, developers and industry, but also for a wider audience.
- making the evaluation methodologies and the results of the data analysis produced in the CLEF campaigns available through the DIRECT system. DIRECT provides the tools for online access of all the information produced during the course of an evaluation campaign, such as runs, topics, relevance assessments, and performance measures, which can be browsed and downloaded for further analyses and re-use.
- distributing the CLEF Evaluation packages through ELDA catalogue. During the CLEF campaigns an infrastructure for evaluating, testing and tuning different types of information retrieval system has been developed. The large variety of data that has been produced not only during the two CLEF campaigns held under TrebleCLEF, but also in CLEF campaign of previous years, has been integrated in the CLEF evaluation packages (document collections, test queries, relevance judgments, and other commonly developed resources such as particular NLP tools etc.) which are documented, packaged, and made available to the MLIA community. ELDA has been in charge of the creation and distribution of such packages, and will continue this dissemination activity also after the end of the project life-time.

As these three activities are pivotal to the Exploitation plan, they are described in detail in the following chapters, explaining the kind of knowledge, data and resources which will be made available, the configuration and management of the infrastructures, the potential target which can be reach and the distribution policy.

2 MLIA Best Practices Portal

The idea behind the MLIA Best Practices Portal is to offer an online repository where different kinds of information, tools and resources related to multilingual information access are made available to the R&D community.

More specifically, the Portal has being originally designed as a major means to further disseminate the Best Practices recommendations about different aspects of the MLIA which have been produced during the project, in order to reach a wider audience through the web. This is a pivotal initiative in the Exploitation Plan, as the Best Practices represent one of the major outcomes of the project and it is considered particularly important to offer an easy modality of access to them beyond the project's lifetime. The Portal has been thus organized mainly around the Best Practices, and all the content has been structured according to a taxonomy of relations starting from four principal nodes corresponding to the Best Practices.

Moving from this primary goal, the Portal has also appeared to be a flexible tool and a valuable resource for sharing a large variety of MLIA-related data and information. In particular, many of the tools and resources that have been produced and used for the development of systems during several years of CLEF campaigns find a natural collocation in the Portal. Some data, such as information about systems that have participated in CLEF QA track, have been already added, and more content taken from the CLEF campaigns will be added in the coming months.

Another important characteristic of Portal is that it has been designed as an interactive tool, where users do not simply search for information, but have also the possibility to actively participate in the

population of the Portal adding new data. The idea is that through the portal all the actors in the field of MLIA -system developers, researchers and also industry- have a virtual space where they can share knowledge and resources. The Portal is not only is a dynamic tool which allows to track different resources for MLIA already available on the web and make them available in a common source similarly to what has been done in other NLP research fields, such as Recognising Textual Entailment (see the Textual Entailment Resource Pool at available at http://www.aclweb.org/aclwiki/index.php?title=Textual_Entailment_Resource_Pool), but also offers the possibility of commenting and giving feedback on the information and data listed, stimulating and enhancing the interaction between all those who operate in the field of MLIA.

Details on the content of the Portal, the modalities of maintaining and updating it, and the plans for future activities are given in the following paragraphs.

2.1 Design and Content of the Portal

The Portal is intended to provide easy access to a variety of MLIA-related resources and tools, organized according to a taxonomy of relations. Currently it is available via the Treble-CLEF website (http://www.trebleclef.eu/jsbestpractices.php), which is due to remain online also after the end of the the project's lifetime.

The menu in the Home Page presents a tree structure, which has four top-level nodes, corresponding to the Best Practices produced in the course of the project, i.e.:

- Evaluation Methodologies (Best Practices for Test Collection Creation, Evaluation Methodologies and Language Processing Technologies
- Language and Evaluation Resources (Best Practices in Language Resources for MLIA
- MLIA System Development System-oriented and User-oriented MLIA Best-practice Recommendations
- Tools (Best Practices in language Resources for MLIA).

The Best Practices in Language Resources has been divided into two different nodes, as tools for the application development have been listed separately, allowing easier search.

Each of these categories has been further divided into sub-categories, where the information taken from the Best Practices and from other MLIA-related sources is displayed in the form of both descriptive texts and links to resources and tool available either inside the Portal or on the web. It is also possible to add HTML fragments and images.

In order to exploit the multilingual aspect of the resources presented, whenever possible indication is given of the language or languages relative to the data presented.

As the portal is intended to be a dynamic resource, it has been designed to be continuously updated and expanded not only by those in charge of the management, but also by users, who after registering can share resources and useful information about MLIA technologies. In this way, from its initial purpose of disseminating the project's Best Practices, the Portal is meant to widen its original scope and become an increasingly comprehensive repository of MLIA resources and tools.

At the moment the content present in the portal has been largely taken from the Best Practices. Other kinds of date, such as documents, information about MLIA system and tool produced during the CLEF campaigns, have already been added –e.g. information about QA systems which have participated in CLEF competitions-, and more will be added in the coming months by CELCT, which will be in charge of the system supervision.

Another important contribution to the population of the portal will be hopefully provided by the users. At the moment, we do not have any data on which to make any estimate, as the Portal was released only at the end of the November, and only those who are interested to add content need to register.

Anyway, the number of visitors and the registrations made in the first two weeks of December (9) is encouraging and confirm that the initiative has raised interest in the community.

2.2 Policy on Access to the Portal

The primary goal of the Portal is to enhance the interaction between actors in the MLIA domain and stimulate them to share common knowledge resources. This is an important service, which on the one hand facilitates those who approach MLIA research and technology for the first time, providing in a single place a variety of useful information, data and resources; on the other hand it also offers an opportunity to maximize the circulation of up-to-date information about the state of the art in MLIA taken from different freely available sources, such as Best Practices, Conference Proceedings and scientific publications. According to this vision, the access to the Portal is and will remain a free. CELCT, which will be in charge of the supervision of the content, will make sure that any copyright issues will be avoided, checking that all the information and data present in the Portal are publicly and freely available.

2.3 Management and Maintenance of the Portal

The Portal has been designed and maintained so far by CELCT. With the consent of the other project partners, CELCT will be in charge of the Portal management also from January 2010. This means that CELCT will take care of the technical infrastructure, and provide assistance to users. Moreover, CELCT will actively participate in the addition of new content to the Portal, carrying out a periodical review of relevant MLIA-related new technologies and publications and adding it to the Portal. Data and resources from previous CLEF campaigns will be also added.

As the portal has been designed as a dynamic repository of information, which is continuously kept up-to date also by the contribution of users that are allowed to add new data, CELCT will also supervise the population activities, selecting and coordinating a pool of experts in the different disciplines covered in the Portal –the authors of the Best Practices in first place-, which will revise the content, assuring a high-qualitative standard of the information present in the different categories.

As long as the TrebleCLEF website will remain online, the portal will be kept as part of it, to emphasise its function of means of dissemination of the results of the project. Thus CELCT will also be in charge of the maintenance of the TrebleCLEF website during 2010.

2.4 Future Dissemination Activities via the Portal

Originally meant to be made public during summer 2009, the Portal was officially published on November 30, 2009. This was mainly due to the fact that it was important to have all the Best Practices represented, the last of which was realised on October 2009.In addition, although a prototype of the Portal was already presented at the Review meeting in January, it took quite a long type to populate it. In fact, the content of the Best Practices had to be selected and organized in a coherent disposition, trying to maximize the link between resources. Moreover, during the population, the partners in charge of the single nodes contributed to test the functionality of the portal, and on the basis of their observations and remarks, some changes were introduced. In any case, the fact that the Best Practices Portal was brought to the public at the end of the TrebleCLEF project, means that it will be a showcase for presenting the results of the many research activities carried out during the project to the MLIA community, and it will be vital to kept it operative and up-to-date after the end of the project. CELCT will be in charge of the maintenance of the infrastructure and supervision of the content at least for all 2010, and possibly afterwards, according to its availability of financial resources and man-power. Also the other members of the consortium –ELDA in particular- will join efforts to keep the Portal alive in the long term perspective, mainly trying to obtain funding from MLIA related projects which can interested in exploiting it for their activities.

According to its original aim, the Portal is mainly targeted to the IR research community and industry, so it is principally to these two targets that the dissemination will be addressed, not excluding a wider

audience –e.g. other different research communities in the NLP areas for which lists of language resources and tools represent a valuable help.

The Portal has been already been advertised through the most relevant mailing list –such as the IR conferences' mailing list (CLEF, TREC, NTCIR, FIRE), but also other NLP conferences (e.g. he Text Analysis Conference and Recognizing Textual Entailment), as well as more general mailing lists such as Corpora and Linguists. Beside explaining the aims and content of the portal, the participation in the portal population has been encouraged. More advertising will be done in the future, sending out periodic (possibly quarterly) summary reports on what is new in the Portal.

An important occasion of dissemination of the Portal will be at CLEF 2010, to be held in Padua in September 2010, due to its strong connections with the previous CLEF campaigns and with the Treble-CLEF project. In fact, researchers and participants in the main conference and in the collocated laboratories will certainly benefit of the resources and tools made available through the Portal. Thus, while promoting the campaign, the participants will be encouraged to actively contribute to the augmentation of knowledge available in the portal, sharing their own resources and tool.

3 DIRECT

As described in [2], the DIRECT system can be used with a twofold aim:

(i) to manage all the steps of an evaluation campaign, i.e. the ingestion of the document collections, the creation of the topics, the submission of the experiments, the relevance assessments, and the computation of performance measures and statistical analyses;

(ii) to make all the managed information online accessible to registered users and to preserve them over the time.

Since the first prototype developed by UNIPD for CLEF 2005, DIRECT has gathered and provides access to: more than 5.6 million documents; more than 1.5 million relevance assessments for more than 600 topics made by over 200 assessors in 15 different countries; more than 2,500 experiments, which amount to about 117 million tuples, submitted by over 170 participants in about 30 different countries; over 5.5 million performance measures and descriptive statistics; and, about 20,000 plots and statistical analyses graphs.

Moreover, since June 2009, an intense activity has been started in order to insert all the data concerning the Ad-hoc, Domain Specific, and GeoCLEF tracks for all the ten years of CLEF from 2000 to 2009 and to harmonize them with the data already managed by the system; this will represent the complete evaluation cycles for the tracks mentioned before. The information gathered through this activity is not limited to the runs, topics, documents, and pools of these ten years of CLEF but it concerns also the papers of the CLEF working notes that describe and explain those experiments and results and which are correctly linked to the corresponding runs and tasks. The outcome of this activity will give users and stakeholders access to an unprecedented amount of online information about experimental evaluation in the MLIA fields as well as links to the papers where the algorithms experimented and the results obtained are explained and detailed. In this way, users will be able not only to access the papers about CLEF (which are already online on the CLEF Web site) but to directly access and explore the described experiments by exploiting the links between the papers and the experiments, consult and download the related performance measures and statistical analyses as well as, given an experiment and all it related information, retrieve the corresponding paper.

At the time of writing, the data ingestion process has been almost completed and UNIPD plans to use the first quarter of 2010 to check and validate the inserted data and to extend the DIRECT systems with the functionalities needed to publish online not only the data themselves but also the papers concerning them. Then, at the end of the first quarter / beginning of the second quarter of 2010, UNIPD will make this information online available to the research and developer communities that will be able to use and re-use them for their own purposes. A first use of these data can be envisioned

with respect to the CLEF 2010 conference, where papers revising and mining ten years of CLEF are solicited.

UNIPD will also make the DIRECT system available to host and manage, upon agreement, the tracks of CLEF 2010 or of other evaluation initiatives which request such support. This will contribute to a wider adoption of DIRECT, to enlarging the knowledge base, to additional development of the systems to tailor it to the specific needs of new tracks and tasks, if needed, and to sustain its development by means of the agreements between UNIPD and the interested partners.

4 CLEF Evaluation Packages

The valuable resources and experimental collections created by the major MLIA evaluation campaigns must be made available for promoting and coordinating the re-use and distribution of these data to the relevant communities. This is an important way to sustain an R&D community by providing high quality access to past evaluation results thus boosting the R&D activities through further dissemination of know-how, tools, resources and best practice guidelines. It is also one of the objectives of Work Package 5 (WP5: *Evaluation Packages and Language Resources*) of the TrebleCLEF project, with respect to the CLEF evaluation resources.

Since the first CLEF campaign in 2000, the CLEF project has been developing an infrastructure for the evaluation, testing and tuning of many different types of information retrieval systems operating on different domains (news, scientific reports, patents, etc.), modalities (text, image, multimodal IR) and languages. The extreme diversity of past CLEF activities and the large amount of data resulting from these evaluation campaigns emphasise the critical need for CLEF evaluation packages (document collections, test queries, relevance judgments, and other commonly developed resources such as particular NLP tools etc.) to be documented, packaged, and made available after the evaluation campaigns.

Two deliverables of Work Package 5 (D5.1.1 [1], and D5.1.2 [4]) gave a complete overview of candidate resources for the design of CLEF evaluation packages, including information about the availability of the corresponding resources for a future distribution.

These deliverables proposed a plan to split the different CLEF resources collected along the years into coherent evaluation packages, and gave an overview of the distribution rights issues for each of them. This was a preparation of the final deliverable of WP5: the CLEF evaluation packages (D5.3), which is due at the end of the project (end of December 2009).

ELDA considers that such packages are very valuables outcomes of the project. It encompasses contributions of a large number of key players and active researchers within NLP and IR communities. ELDA has elaborated its exploitation plan, bearing in mind the possibility for a wider community to have access to such resources free of charge. ELDA does not expect any direct Financial Return On Investment from the distribution of such packages but rather considers its involvement in such a project as a service to the community.

4.1 Design of CLEF Packages

The evaluation package of a CLEF task contains the following resources:

- A document describing in detail the content of the package, as well as the corresponding evaluation (tasks, metrics, participants, results, etc.),
- The development and test data collections (sets of documents) and the corresponding annotations when necessary,
- The corresponding sets of topics and queries.

In addition, a range of specific data is provided for each evaluation task, allowing the package user to reproduce the evaluation in conditions “similar” to the evaluation campaign and to compare their results with those of the participants:

- Documentation about the evaluation procedure (evaluation tools, submission format, etc.),
- The input data, as received by the participants during the evaluation,
- The evaluation and scoring tools,
- The participants' submissions and results (provided that the concerned task's participants gave their agreement).

The test-suites will also contain the reports prepared by the participating groups describing their experiments, in order to ensure repeatability. These are the CLEF Working Notes, produced each year for the Workshop.

The content of test suites are validated by the project partners and participants, who provided baseline reference results, while testing the technologies which are part of their recognized domain of expertise.

4.2 Evaluation Task Coverage

A first CLEF evaluation package was already produced within the first years of CLEF. It is distributed by ELDA through the ELDA/ELRA catalogue[2]. This package covers CLEF campaigns between 2000 and 2003 (i.e. the four first years of CLEF).

The goal of TrebleCLEF is to cover the years 2004-2009 , making as many CLEF resources as possible publicly available to the academic and industrial communities as a set of distinct *task evaluation packages*.

ELDA is presently working on the constitution, quality validation and distribution of the first CLEF evaluation packages. The work plan for the end of the project is:

- November and December 2009: Finalisation of the "core" set of CLEF evaluation packages, which concern the following tracks:
  - AdHoc on News data (2004-2008),
  - QA@CLEF (2003-2008),
  - Domain Specific (2004-2008),
  - GeoCLEF (2005-2008).
- December 31st 2009: Final delivery of the above CLEF packages (D5.3).
- January 2010: Integration into the ELDA catalogue.
- First quarter of 2010: Examine the possible integration of further CLEF packages into the ELDA catalogue (depending on which negotiations could be finalized).

For now we will not consider the integration of data after 2008 (i.e. 2009), since it has only been possible to negotiate the use of data up to 2008 for the moment. Data of the 2009 campaign may be considered after the project, to form new packages or enrich the existing ones.

4.3 Negotiation of Intellectual Property Rights

The question of Intellectual Property Rights (IPR) is a key issue for the distribution of the packaged evaluation resources. ELDA is in charge of negotiating with IPR owners (i.e. content providers and tool developers) to ensure that in all cases their rights will be respected. Agreements have to be drafted according to the laws currently in force, considering the concerns and the rights of the resources providers and the expectations of the users.

4.4 Distribution

ELDA has always endeavoured to distribute evaluation packages at the lowest possible price (despite the production and maintenance costs), in order to guarantee fair access to resources to the widest range of HLT actors. The price of the package is calculated as the fee requested by the data providers

[2] ELDA/ELRA catalogue: http://catalog.elra.info/

(where applicable) plus ELDA's nominal distribution charge. ELDA seeks to negotiate the lowest possible fee with the data providers. Negotiations will be started with the CLEF data providers.

In the past years, most data providers have not sought any financial compensation for the distribution of their data in evaluation packages. The best case scenario envisaged is that ELDA can distribute the evaluation package at its nominal distribution charge i.e. 'at shipment cost' (often 0€ when this is doable via Internet).

ELDA is thus working on securing a fair distribution procedure on behalf of the CLEF consortium, namely:

- Establishment of a simplified licensing scheme, mostly through the negotiation of distribution rights for the text data collection, via the exploitation of existing, ready-to-use distribution and end-user agreements;
- Exploitation of a regular distribution channel, such as the existing online ELDA/ELRA resource catalogue[3]. Resource distribution is supported by proven promotion techniques, already implemented for ELDA's everyday activity.

4.5 Dissemination of Evaluation Package

ELRA/ELDA will make full use of its dissemination channels (ELRA newsletter, LREC conference, mailing lists, etc.) in order to promote awareness of the availability of the reusable language resources (the test collection of multilingual language resources for evaluation purposes) and the evaluation procedures and methodologies created by CLEF. The goal will be to recoup the investment of much effort made by a large number of research groups and to avoid duplication of efforts by other researchers and other projects. This dissemination activity should boost the evaluation paradigm in Europe and will certainly contribute to the enhancement of take-ups and demonstration activities that require Language Resources for training and development.

4.6 Reusability

Although the activities of CLEF have focussed on the needs of multilingual information retrieval system developers, the language resources produced will also be useful to other technology developers, service providers, and corporate end-users. The availability of such resources will support technology players in developing new applications or in customising/porting existing ones onto different languages, domains or user groups. We expect that the need for transfer of innovative technologies to a wider set of languages would encourage a number of HLT players to acquire the test suites.

The strengthening and widening of the access to information resources for a wider range of languages and language groups is a major point in the EU policies on e-commerce and cultural heritage. The results of CLEF can be exploited in a number of important European application areas in which multicultural and multilingual issues play a key role: distance learning and education; on-line access to sources of cultural heritage, such as archives, museums, digital libraries, etc.; e-commerce; international law archives; European laws and acts of parliament, etc.

5 Concluding Remarks

Our aim in the exploitation plan is to ensure that the results of TrebleCLEF can not only be disseminated widely but can also be built on and extended in the future. The three tools for exploitation described in this report represent diverse ways in which the interested research and user communities can access and reuse the collection of valuable data, information and knowledge created within the project and are directed at different target groups. Taken together, they are intended to

[3] ELDA Catalogue of Language Resources: http://catalog.elra.info/

provide a complementary and exhaustive coverage and dissemination of the tools, resources and knowhow inherent in the Multilingual Information Access (MLIA) domain:

- The Portal will provide the active MLIA community of researchers and developers and interested newcomers with a central reference point for access to information on the best practices and the tools and resources available for this domain; it is intended as a dynamic and updatable networking tool for all kinds of users.
- The evaluation packages will release the test suites produced in the diverse evaluation tracks and campaigns in a reusable form for system benchmarking and experimental purposes; they make it possible for developers to run testing and tuning experiments according to their own needs, independently of the timing and schedule of an evaluation campaign.
- DIRECT will provide not only the data of the evaluation packages but additional tools to process and manipulate the data thus supporting further in-depth experimental studies with the aim of facilitating research advances in the MLIA domain.

We will strive to exploit these connections as much as possible to promote the dissemination and use of these outcomes of TrebleCLEF and to let potential users understand which is the best tool for their needs. Each tool will provide explicit links to the others, for example, links can be added to the description of resources in the TrebleCLEF Portal to directly point to relevant experiments and tasks in DIRECT.

6 References

[1] Dussin, M. and Ferro, N., Managing the Knowledge Creation Process of Large-Scale Evaluation Campaigns. In M. Agosti, J. Borbinha, S. Kapidakis, C. Papatheodorou, G. Tsakonas, editors, *Proc. 13th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2009)*, pages 63–74. Lecture Notes in Computer Science (LNCS) 5714, Springer, Heidelberg, Germany, 2009.

[2] Ferro, N., *Operational Scientific Digital Library*, Public report of the TrebleCLEF project (Deliverable 2.2), July 2008.

[3] Moreau N. *et al.*, *Evaluation Resources for CLEF*, Public report of the TrebleCLEF project (Deliverable 5.1.1), September 2008.

[4] Moreau N. *et al.*, *Evaluation Resources for CLEF*, Public report of the TrebleCLEF project (Deliverable 5.1.2), September 2009.