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 D5.3
CLEF EVALUATION PACKAGES

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

ELDA, ISTI-CNR, CELCT, USFD

Version: final

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

---

Document Information

Deliverable number:	5.3
Deliverable title:	CLEF Evaluation Packages
Due date of deliverable:	12|2009
Actual date of deliverable:	12|2009
Author(s):	Nicolas Moreau, ELDA
Participant(s):	All
Workpackage:	5
Workpackage title:	Evaluation Packages and Language Resources
Workpackage leader:	ELDA
Est. person months:	3.5
Dissemination Level:	PU
Version:	V2.0
Keywords:	Language Resources, Data Collection, Evaluation Packages

History of Versions

Version	Date	Status	Author (Partner)	Description/Approval Level
V1.0	15/12/2009	Preliminary	ELDA (N. Moreau)	1rst circulated preliminary version
V1.1	22/12/2009	Nr-Final	ELDA (N. Moreau)	Comments from partners
V1.2	23/12/2009	Pre-Final	ELDA (N. Moreau)	
V2.0	31/12/2009	Final	ELDA (N. Moreau)	

Abstract

The purpose of this deliverable is to describe the test collections created within the framework of the CLEF evaluation campaigns that have been packaged into test-suites. These test suites will be distributed through the ELRA/ELDA catalogue. They can be used for system benchmarking and other R&D activities in the Information Retrieval and Human Language Technology sector.

This is a key action in the dissemination of resources and methodologies implemented by CLEF for the testing and tuning of information retrieval systems working in the multilingual context.

Table of Contents

Executive Summary

The main goal of WP5 has been to design publicly available test-suites containing all the necessary data that can be re-used outside the evaluation campaigns for system benchmarking. WP5 had to ensure that these resources (data, metrics, reports, methodologies) are made widely available and licensed with a clean IPR and copyright agreements. The CLEF resources will be made available to the community through the constitution of *CLEF evaluation packages*.

The question of Intellectual Property Rights (IPR) is an issue of importance for the packaging of evaluation resources. WP5 negotiated with IPR owners (i.e. content providers and tool developers) to ensure that in all cases their rights will be respected. Agreements were formulated according to the laws currently in force, and both with reference to content provision (agreements with content providers) and content exploitation (agreements with end users).

This deliverable describes how part of the CLEF resources collected along the years were split into coherent evaluation packages, and gives an overview of each of them. This is the final deliverable of WP5.

1 Introduction

The valuable resources and experimental collections created by the major MLIA evaluation campaigns must be made available by promoting and coordinating the re-use and distribution of these data to the relevant communities [1]. 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. This is one of the objectives 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 resulting data from these evaluation campaigns emphasises 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 designed and made available after the evaluation campaigns.

The first Deliverable of Work Package 5 (Deliverable D5.1.1 [2]) gave an initial overview of candidate resources for the design of CLEF 2008 packages.

The third WP5 Deliverable (Deliverable D5.1.2 [3]) extended this overview to all the CLEF campaign years since CLEF begun (in 2000) and gave additional information about the availability of resources.

This final WP5 Deliverable (D5.3) describes the CLEF evaluation packages that are to be released within TrebleCLEF as well as ELDA's distribution policy. It also gives some information on additional packages that can be prepared after the end of the project as soon as remaining copyright problems have been cleared.

2 Overview of CLEF Tasks and Packages

The tasks and subtasks addressed by the released CLEF packages are displayed in Table 1:

Task	Sub-task	2000	2001	2002	2003	2004	2005	2006	2007	2008
AdHoc	CLIR on News	•	•	•	•	•	AdHoc News 2004-2008			•
AdHoc	Robust	CLEF 2000-2003						•	•	•
AdHoc	TEL@CLEF									•
Domain specific	Scientific data retrieval	•	•	•	•	•	Domain Specific 2004-2008			•
QA@CLEF	Main QA				•	•	QA 2003-2008		•	•
QA@CLEF	QAST (Speech Transcripts)								•	•
GeoCLEF	Main						•	•	•	•
INFILE@CLEF	INFILE@CLEF									•

Table 1: Released CLEF packages.

The released packages are:

- CLEF 2000-2003 (released before the start of TrebleCLEF).
- AdHoc News 2004-2008: mono- and cross-language search on newspaper corpora.
- Domain Specific 2004-2008: mono- and cross-language search on domain-specific data (mainly social sciences articles).
- QA 2003-2008: mono- and cross-language question answering (news articles and Wikipedia).

The release of additional CLEF packages can be envisaged after the end of the project depending on the conclusion of rights negotiations:

- AdHoc - Robust 2006-2008: retrieval tasks on newspaper corpora based on the use of word sense disambiguated (WSD) data.
- AdHoc - TEL@CLEF: mono- and cross-language search on library catalogue records, organized in collaboration with The European Library (TEL)[1].
- QAST: question answering on speech transcriptions of seminars.
- GeoCLEF 2003-2008: cross-language geographic information retrieval.
- INFILE: mono- and cross-language filtering evaluation.

3 Released CLEF Packages

The CLEF campaigns from 2000 to 2009 offered many different kinds of tracks, evaluation sub-tasks and resources. It was not possible to address all data within the 2 years of TrebleCLEF and the priority was given to some core tracks of CLEF (AdHoc, Domain Specific , QA).

The CLEF test collections are basically made up of documents, topics and relevance assessments. An evaluation package consists of the test collection(s) with full documentation (including definition and description of the evaluation methodologies, protocols, and metrics), along with the software scoring tools that are necessary to evaluate developed systems for a given technology. Such a package provides system developers that did not participate in the evaluation campaigns with the necessary tools to benchmark their systems and compare results to those obtained during the official evaluation.

The following section describes the CLEF evaluation packages released by the end of TrebleCLEF.

Here, we will not discuss 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 and later may be considered after the project, to form new packages or enrich the existing ones.

3.1 Structure of Evaluation Packages

The structure of new packages follows the HTML structure of the previously released CLEF2000-2003 evaluation package. A CLEF evaluation package consists of:

- An introduction page introducing and giving access to the content of the package.
- General information on the documentation about the evaluation procedure (evaluation tools, submission format, etc.),
- Guidelines for each year's campaign (based on the information provided to the participants during the actual official campaign) about the evaluation procedure: data format (input), submission format (output), etc.
- The data collections (sets of documents).
- The sets of topics or questions.
- The scoring tools.
- A description of how to reproduce the evaluation with one's own system.
- The official participants' submissions and relevant assessments.
- The official results of the different campaigns

The test-suite 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 was validated by the project partners and task coordinators.

[1] See http://www.theeuropeanlibrary.org/

3.2 CLEF AdHoc – News 2004-2008

This package covers the AdHoc tracks on news data after 2003 until 2008. It comes in continuity with the CLEF 2000-2003 Package that had already been produced within the first years of CLEF. It has been distributed for a few years by ELDA through the ELDA/ELRA catalogue[1]. This CLEF 2000-2003 package covers CLEF AdHoc and Domain Specific campaigns between 2000 and 2003 (i.e. the four first years of CLEF).

The overall multilingual news corpus contains more than 3 million news documents in 14 European languages. This corpus is divided into two comparable collections (i.e. addressing comparable issues, since the news documents included were published during the same time period):

- 1994-1995: Dutch, English, Finnish, French, German, Italian, Portuguese, Russian, Spanish, Swedish,
- 2000-2002: Basque, Bulgarian, Czech, English, Greek, Hungarian.

In addition, a collection covering a non European language (Persian) was used by the Ad Hoc track in 2008: the Hamshahri Persian newspaper corpus; nearly 170,000 documents. These data collections are detailed in Appendix A. What datasets are to be used for each year's campaign is described in Appendix B.

3.3 QA@CLEF - News 2003-2008

This package covers the main QA track dealing with news data from 2003 to 2008. There was a QA pilot track in 2003, but it was not covered by the CLEF 2000-2003 Package and hence is included in this new one.

Basically, the news datasets are the same as the ones used for the AdHoc main track during that period (Appendix A). Wikipedia page dumps (of November 2006) were also used in the later campaigns (QA 2006 – 2008) and can be downloaded from the web: the required web links are provided in the package. What data collections were used for the different years of the QA evaluation campaigns are detailed in Appendix C:

In the main 2009 QA campaign (called ResPubliQA[2]), a completely different multilingual parallel dataset was used. Since this is a different sort of data and tasks, it could be released in 2010 in another distinct package if ELDA is able to stipulate a distribution agreement with the providers.

3.4 CLEF Domain Specific 2004-2008

This package covers the Domain Specific tracks after 2003 and until 2008 (in continuity to the CLEF 2000-2003 Package).

The main datasets were GIRT-4, a social science database in English and German (over 300,000 documents) and two Russian databases: the Russian Social Science Corpus (approx. 95,000 documents) and the Russian ISISS collection for sociology and economics (approx. 150,000 docs). The Cambridge Sociological Abstracts corpus in English (20,000 docs) was later introduced. All data collections used in the CLEF Domain-Specific track are detailed in Appendix D:

- Domain Specific are listed in section D.1
- How Domain Specific corpora were used in the different campaigns is detailed in the table of section D.2.

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

[2] ResPubliQA: http://celct.isti.cnr.it/ResPubliQA/

4 Distribution and Dissemination

4.1 Distribution

ELDA has always endeavoured to produce and distribute evaluation packages at the lowest possible price, in order to guarantee fair access to resources to a range of HLT actors. The price of the package is calculated as the fee requested by the data providers (where applicable) plus ELDA's nominal distribution charge. ELDA seeks to negotiate the lowest possible fee with the data providers.

Overall, ELDA's pricing policy regarding evaluation packages distinguishes between different types of use (research and commercial use) according to the type of user organisation (academic or commercial organisation). Moreover, there are two distinct price blocks, one dedicated to ELRA Members and one to non Members.

For the new CLEF test suites, ELDA will distribute the evaluation packages at their nominal distribution charge i.e. 'at cost' for academic organisations.

For commercial organisations the price will be approximately the same as for the CLEF 2000-2003 package, i.e.: around EUR 500.00 for ELRA members, and around EUR 1,000.00 for non members.

For distribution of such packages, ELDA uses its regular distribution channel, i.e. the online ELDA/ELRA resource catalogu[1], as it was already done for the first CLEF 2000-2003 package. Resource distribution is supported by proven promotion techniques, already implemented for ELDA's everyday activity.

4.2 Dissemination

The TrebleCLEF consortium is aware of the importance of disseminating the results achieved by CLEF to as wide as possible potentially interested R&D and application communities. ELRA/ELDA is naturally bound to play a large role in this activity as the agency has developed a range of communication channels to promote language resource and evaluation activities in HLT. These include:

- The ELRA Newsletter, distributed to over a thousand HLT actors worldwide, including a special issue dedicated to CLEF;
- The high-level conference on Language Resources and Evaluation (LREC), that can boast hundreds of participants from HLT key sectors and offers CLEF one of the most important showcases in the HLT community;
- An active website, with pages promoting evaluation activity in CLEF;
- Daily monitoring and exploitation of international HLT mailing lists.

ELRA/ELDA will make full use of its dissemination channels 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 IST 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.

5 Conclusion and Additional Packages

This deliverable has given a final overview of evaluation resources packaged within the TrebleCLEF project.

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

Following the first released CLEF evaluation package, covering the first years of CLEF (2000-2003), the TrebleCLEF project releases 3 new CLEF test suites covering the following years till 2008.

A first set of core CLEF tracks, have been packaged, covering the 3 main historical core tracks that are: AdHoc, Domain Specific and Question Answering.

For a second set of tracks, a few minor remaining ownership issues are being negotiated. Once all copyright problems are solved for all or some of these tasks, ELDA will finalise the preparation of additional test suites, to be released in 2010. As mentioned before in section 2, the potential set of additional tasks is:

- AdHoc - Robust.
- AdHoc – TEL.
- QAST.
- GeoCLEF.
- INFILE.

A third set of resources (in particular resources for the ImageCLEF and VideoCLEF tracks) could not be addressed. They still require of lot of effort in terms of technical validation and IPR negotiations before they can be distributed properly.

ELDA has worked out a legal framework for the consortium to have access to the test collections and a licensing schema that will ensure a wider availability of CLEF packages.

In order to make sure the released (and upcoming) evaluation packages are distributed outside the consortium all legal rights, licensing schemas, commercial and pricing issues have been (or will be) addressed and cleared.

The distribution of the CLEF evaluation packages will be done via the ELRA/ELDA resource catalogue.

As for any new resource included in the catalogue, released CLEF packages will be advertised through many different channels, in particular the ELRA newsletters.

6 References

[1] Moreau N. *et al., Best Practices in Language Resources for Multilingual Information Access*, Public report of the TrebleCLEF project (Deliverable 5.2), March 2009.

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

[3] Moreau N. *et al., CLEF Evaluation Resources*, Public report of the TrebleCLEF project (Deliverable 5.1.2), September 2008.

Appendix A CLEF News Datasets

The following tables describe the news datasets (newswire and newspaper articles) used in different CLEF evaluation campaigns.

Resource Name	Description	Provider
Egunkaria 2000-02 (Basque)	Egunkaria newspaper, 2000-2002 918 XML files, 119,982 documents, 212 MB	Egunkaria
Sega 2002 (Bulgarian)	Sega newspaper, 2002 33,356 documents, 120 MB	Sega AD, Newspaper Sega, Sofia, Bulgaria
Standart 2002 (Bulgarian)	Standart newspaper, 2002 35,839 documents, 93 MB	Standart NYUZ AD, Newspaper Standart, Sofia, Bulgaria
Novinar 2002 (Bulgarian)	Novinar newspaper, 2002 18,086 documents, 48 MB	Novinar OOD, Newspaper Novinar, Sofia, Bulgaria
Mlada fronta 2002 (Czech)	Mlada fronta DNES newspaper, 2002 68,842 documents, 143 MB	Mafra a.s
Lidové Noviny 2002 (Czech)	Lidové Noviny newspaper, 2002 12,893 documents, 35 MB	Lidové Noviny
NRC Handelsblad 94-95 (Dutch)	NRC Handelsblad newspaper, 1994-1995 84,121 documents, 299 MB	PCM Landelijke dagbladen/Het Parool.
Algemeen Dagblad 94-95 (Dutch)	Algemeen Dagblad newspaper, 1994-1995 106,483 documents, 241 MB	PCM Landelijke dagbladen/Het Parool.
LA Times 94 (English)	Los Angeles Times newspaper, 1994 113,005 documents, 425 MB	LA Times
LA Times 2002 (English)	Los Angeles Times newspaper, 2002 135,153 documents, 434 MB	LA Times
Glasgow Herald 95 (English)	Glasgow Herald newspaper, 1995 56,472 documents, 154 MB	Newquest (Herald & Times) Ltd.
Aamulehti 94-95 (Finnish)	Aamulehti newspaper, 1994-1995 55,344 documents, 137 MB	Aamulehti
Le Monde 94 (French)	Le Monde newspaper, 1994 44,013 documents, 157 MB	Le Monde
Le Monde 95 (French)	Le Monde newspaper, 1995 47,646 documents, 156 MB	Le Monde
SDA 94 (French)	SDA - French, 1994 43,178 documents, 86 MB	Schweizerische Depeschenagentur AG
SDA 95 (French)	SDA - French, 1995 42,615 documents, 88 MB	Schweizerische Depeschenagentur AG
Frankfurter Rundschau 94 (German)	Frankfurter Rundschau newspaper, 1994 139,715 documents, 320 MB	Der Spiegel
Der Spiegel 94-95 (German)	Der Spiegel newspaper, 1994-1995 13,979 documents, 63 MB	Der Spiegel
SDA 94 (German)	SDA - German, 1994 71,677 documents, 144 MB	Schweizerische Depeschenagentur AG
SDA 95 (German)	SDA - German, 1995 69,438 documents, 141 MB	Schweizerische Depeschenagentur AG
S-E European Times 2002 (Greek)	The Southeast European Times newspaper, 2002	S-E European Times
Magyar Hirlap 2002 (Hungarian)	Magyar Hirlap newspaper, 2002 49,530 documents, 105 MB	A-PONT-MH Publisher Ltd

Resource Name	Description	Provider
La Stampa 94 (Italian)	La Stampa newspaper, 1994 58,051 documents, 193 MB	La Stampa
SDA 94 (Italian)	SDA - Italian, 1994 50,527 documents, 85 MB	Schweizerische Depeschenagentur AG
SDA 95 (Italian)	SDA - Italian, 1995 48,980 documents, 85 MB	Schweizerische Depeschenagentur AG
Hamshahri 1996-2002 (Persian)	Hamshahri newspaper, 1996-2002 166,774 documents, 611 MB	DBRG-University of Tehran
Público 94 (Portuguese)	Público newspaper, 1994 51,751 docs, 164 MB	Público
Público 95 (Portuguese)	Público newspaper, 1995 55,070 docs, 176 MB	Público
Folha de São Paulo 94 (Portuguese)	Folha de São Paulo newspaper, 1994 51,875 documents, 108 MB	Folha de São Paulo
Folha de São Paulo 95 (Portuguese)	Folha de São Paulo newspaper, 1995 52,038 documents, 116 MB	Folha de São Paulo
Izvestia 95 (Russian)	Izvestia newspaper, 1995 16,761 documents, 68 MB	Izvestia
EFE 94 (Spanish)	EFE news agency, 1994 215,738 documents, 509 MB	Agencia EFE S.A.
EFE 95 (Spanish)	EFE news agency, 1995 238,307 documents, 577 MB	Agencia EFE S.A.
TT 94-95 (Swedish)	Tidningarnas Telegrambyrå newspaper, 94-95 142,819 documents, 352 MB	TT

Appendix B AdHoc News 2004-2008: Datasets

	News Datasets - CLEF AdHoc																																	
	BG		CS		DE				EN			ES		FA	FI	FR				HU	IT			NL		PT				RU	SV	
Tasks	Sega 2002	Standart 2002	Lidové Noviny 2002	Mlada Fronta 2002	Frankfurter Rundschau 94	Der Spiegel 94-95	SDA 94	SDA 95	LA Times 94	LA Times 2002	Glasgow Herald 95	EFE 94	EFE 95	Hamshahri, 1996-2002	Aamulehti Nov.94-95	Le Monde 94	Le Monde 95	SDA 94	SDA 95	Magyar Hirlap 2002	La Stampa 94	SDA 94	SDA 95	NRC Handelsblad 94-95	Algemeen Dagblad 94-95	Publico 94	Publico 95	Folha de São Paulo 94	Folha de São Paulo 95	Izvestia 95	TT 94-95	Tasks
AdHoc 2004																																		**AdHoc 2004**
Multilingual											X				X		X		X											X				Multilingual
Bilingual											X				X		X		X								X			X				Bilingual
Monolingual															X		X		X								X			X				Monolingual
AdHoc 2005																																		**AdHoc 2005**
Multilingual					X	X	X	X	X		X	X	X		X	X		X	X		X	X	X	X	X								X	Multilingual
Bilingual	X	X							X		X					X	X	X	X	X						X	X	X	X					Bilingual
Monolingual	X	X														X	X	X	X	X						X	X	X	X					Monolingual
AdHoc 2006																																		**AdHoc 2006**
Bilingual	X	X							X		X					X	X	X	X	X						X	X	X	X					Bilingual
Monolingual	X	X														X	X	X	X	X						X	X	X	X					Monolingual
AdHoc 2007																																		**AdHoc 2007**
Bilingual	X	X	X	X						X										X														Bilingual
Monolingual	X	X	X	X																X														Monolingual
Indian										X																								Indian
AdHoc 2008																																		**AdHoc 2008**
Persian														X																				Persian

Appendix C QA 2003-2008: Datasets

QA@CLEF - Test Collections

	News Datasets																															Wikipedia (Nov. 2006)												
	BG			DE				EL	EN			ES		EU	FI	FR				IT			NL		PT			RO	BG	DE	EL	EN	ES	EU	FI	FR	IT	NL	PT	RO	
Tasks	Sega 2002	Standart 2002	Novinar 2002	Frankfurter Rundschau 94	Der Spiegel 94-95	SDA 94	SDA 95	SE European Times 2002	LA Times 94	LA Times 2002	Glasgow Herald 95	EFE 94	EFE 95	Egunkaria, 2000-2002	Aamulehti Nov.94-95	Le Monde 94	Le Monde 95	SDA 94	SDA 95	La Stampa 94	SDA 94	SDA 95	NRC Handelsblad 94-95	Algemeen Dagblad 94-95	Publico 94	Publico 95	Folha de São Paulo 94-95	*no news data set*	Wikipedia Bulgarian	Wikipedia German	Wikipedia Greek	Wikipedia English	Wikipedia Spanish	Wikipedia Basque	Wikipedia Finnish	Wikipedia French	Wikipedia Italian	Wikipedia Dutch	Wikipedia Portuguese	Wikipedia Romanian	Tasks
QA 2003																																										**QA 2003**
Monolingual												X								X	X		X	X																		Monolingual
Crosslingual									X																																	Crosslingual
QA 2004																																										**QA 2004**
Monolingual				X	X	X	X					X	X			X		X	X	X	X	X	X	X	X	X																Monolingual
Crosslingual				X	X	X	X		X		X	X	X			X		X	X	X	X	X	X	X	X	X																Crosslingual
QA 2005																																										**QA 2005**
Monolingual	X	X		X	X	X	X					X	X		X	X		X	X	X	X	X	X	X	X	X	X															Monolingual
Crosslingual	X	X		X	X	X	X		X		X	X	X		X	X		X	X	X	X	X	X	X	X	X	X															Crosslingual
QA 2006																																											**QA 2006**
Monolingual	X	X		X	X	X	X					X	X			X	X	X	X	X	X	X	X	X	X	X	X															Monolingual
Crosslingual	X	X		X	X	X	X		X		X	X	X			X	X	X	X	X	X	X	X	X	X	X	X															Crosslingual
QA 2007																																											**QA 2007**
Monolingual	X	X	X	X	X	X	X					X	X			X	X	X	X	X	X	X	X	X	X	X	X		X	X			X			X	X	X	X	X	Monolingual
Crosslingual	X	X	X	X	X	X	X		X		X	X	X			X	X	X	X	X	X	X	X	X	X	X	X		X	X		X	X			X	X	X	X	X	Crosslingual
QA 2008																																											**QA 2008**
Monolingual	X	X	X	X	X	X	X	X				X	X	X		X	X	X	X	X	X	X	X	X	X	X	X		X	X	X		X	X		X	X	X	X	X	Monolingual
Crosslingual	X	X	X	X	X	X	X		X		X	X	X	X		X	X	X	X	X	X	X	X	X	X	X	X		X	X		X	X	X		X	X	X	X	X	Crosslingual

Appendix D CLEF Domain Specific

D.1 Datasets Description

Resource Name	Description	Provider
GIRT-4	Collection of German social science data. 2 parallel corpora containing the same 151,319 documents: · German GIRT4 (GIRT4-DE) · English GIRT4 (GIRT4-EN) Size: 302,638 docs, 524 MB	GESIS (Informationszentrum Sozialwissenschaften der Arbeitsgemeinschaft Sozialwissenschaftlicher Institute e.V.)
RSSC	Russian sociology database data from the Russian Social Science Corpus 94,581 docs, 65 MB	ISIS RAS (Institute of Scientific Information for Social Sciences of the Russian Academy of Science)
INION-ISISS	The INION-ISISS corpus covers the Russian social sciences and economics. 145,802 docs, 12 MB	ISIS RAS (Institute of Scientific Information for Social Sciences of the Russian Academy of Science)
CSA-SA	Database of Sociological Abstracts from Cambridge Scientific Abstracts (CSA) 20,000 docs, 38.5 MB	Proquest CSA (Cambridge Information Group)

D.2 Domain Specific 2004-2008: Datasets

	CLEF Domain Specific - Target Collections					
	DE	EN		RU		
Campaign	GIRT4	GIRT4	CSA-SA	RSCC	INION	Campaign
CLEF 2004	X	X				CLEF 2004
CLEF 2005	X	X		X		CLEF 2005
CLEF 2006	X	X		X	X	CLEF 2006
CLEF 2007	X	X	X		X	CLEF 2007
CLEF 2008	X	X	X		X	CLEF 2008
Campaign	GIRT4	GIRT4	CSA-SA	RSCC	INION	Campaign
	DE	EN		RU		
	CLEF Domain Specific - Target Collections					