EUROPEAN COMMISSION

7th Framework Programme for
Research, technological
Development and Demonstration

# Table of contents

EUROPEAN COMMISSION

7th Framework Programme for
Research, technological
Development and Demonstration

# A1: Content

| | | | |
|---|---|---|---|
| Proposal Number | 610864 | Proposal Acronym | SEMACORE |
| Project Type | CP-FP-INFSO | | |

## General Information

Proposal Title: SEmantic Multimedia Analysis performed CrOss-language in near REaltime

*Note that for technical reasons, the following characters are not accepted in the Proposal Title and will be removed: < > " &*

| | | | |
|---|---|---|---|
| Duration in months | 36 | Call (part) Identifier | FP7-ICT-2013-10 |

Activity code(s) most relevant to your topic: ICT-2013.4.1

Abstract (max. 2000 chars)

The widespread adoption of smartphones and social media and the availability of Big Data processing capabilities for the first time facilitate the comprehensive analysis of citizen's statements and communication. In particular a multilingual evaluation of the contents of web documents has become possible, offering huge potential societal and economic benefits. SEMACORE will advance the semantic characterization of the web in three crucial aspects: (1) develop methods to analyse content and opinions in multiple languages with uniform semantic definitions, (2) simultaneous analysis of contents and opinions in text and speech from audio/video (3) near real-time exploration of contents in multiple languages and media to cover current events.
The goal of the SEMACORE project is to radically advance the analysis of cross-lingual contents and opinions by developing innovative technologies, methodologies and systems that will put new capabilities in the hand of political and economic decision makers. This will offer a joint and dynamic perception of the opinions of European citizens.
SEMACORE brings together a strong group of researchers with domain experts in three representative use cases on the joint analysis of language, content, and opinions in web documents. The World Network for Linguistic Diversity MAAYA will analyse language use and contents in small countries, international organizations and digital libraries. Nielsen, a leading global media and market research company, will evaluate contents and opinions for an international organization and a company. The technology on speech recognition, machine translation and multilingual content/opinion mining will be provided and developed further by CNRS, UPC, Fraunhofer, and CWI. DIALOGIC will monitor a representative volunteer panel of PC and smartphone users. SEMACORE will provide a standardized workflow for including new languages in the joint analysis opening outstanding avenues for exploitation in business and policy

EUROPEAN COMMISSION

7th Framework Programme for
Research, technological
Development and Demonstration

Free keywords

multimedia, sentiment analysis, language diversity

a) Has this proposal (or a very similar one) been previously submitted to a call for proposals of the 7th EU RTD Framework Programme ? ○ Yes ◉ No

b) Is this proposal (or a similar one) currently being submitted to another call under FP7 ? ○ Yes ◉ No

EUROPEAN COMMISSION

7th Framework Programme for
Research, technological
Development and Demonstration

# A2.1 Participant #1

## ERCIM

If your organisation has already registered for FP7, enter your Participant Identity Code ***999593828***

Legal Name ***GEIE ERCIM***

Organisation short name ***ERCIM***

## Administrative data (legal address)

Street name ***2004, ROUTE DES LUCIOLES SOPHIA ANTIPOLIS*** Number

Town ***BIOT***

Postal Code / Cedex ***06410***

Country ***FR***

Internet homepage ***http://www.ercim.org***

## Status of your organisation

Certain types of organisations benefit from special conditions under the FP7 participant rules. The Commission also collects data for statistical purposes.

The guidance notes will help you complete this section.

Please 'tick' the relevant box(es) if your organisation falls into one ore more of the following categories.

| | | |
|---|---|---|
| Non-profit organisation | ○ Yes | ◉ No |
| Public body | ○ Yes | ◉ No |
| Research organisation | ○ Yes | ◉ No |
| Higher or secondary education establishment | ○ Yes | ◉ No |

Main area of activity (NACE code) 93

EUROPEAN COMMISSION

7th Framework Programme for
Research, technological
Development and Demonstration

| | | |
|---|---|---|
| 1. Is your number of employees smaller than 250? (full time equivalent) | (•) Yes | ( ) No |
| 2. Is your annual turnover smaller than € 50 million? | (•) Yes | ( ) No |
| 3. Is your annual balance sheet total smaller than € 43 million? | (•) Yes | ( ) No |
| 4. Are you an autonomous legal entity? | (•) Yes | ( ) No |

You are NOT an SME if your answer to question 1 is "NO" and/or your answer to both questions 2 and 3 is "NO".

In all other cases, you might conform to the Commission's definition of an SME.

Please check the additional conditions given in the guidance notes to the forms.

Following this check, do you conform to the Commission's definition of an SME? (•) Yes ( ) No

# Dependencies with (an)other participant(s)

Are there dependencies between your organisation and (an)other participant(s) in this proposal? ( ) Yes (•) No

EUROPEAN COMMISSION

7th Framework Programme for
Research, technological
Development and Demonstration

# Contact point - Person in charge for participant #1

For the co-ordinator (Participant #1) this person is the one who the Commission will contact in the first instance.

| | | | |
|---|---|---|---|
| Family name* | ***Rohou*** | First name(s)* | ***Philippe*** |
| Title | Mr. | ◉ Male | ○ Female |
| Position in the organisation | European project coodinator | | |
| Department/Faculty/Institute/Laboratory name/... | ERCIM Office | | |

**Address** ☒ Same as legal address

| | | | |
|---|---|---|---|
| Street name | 2004, ROUTE DES LUCIOLES SOPHIA ANTIPOLIS | Number | |
| Town | BIOT | Postal Code/Cedex | 06410 |
| Country | FR | | |

| | | | | | |
|---|---|---|---|---|---|
| Phone1* | + | | Phone2 | + 33 | 628474072 |
| Fax | + 33 | 492387822 | E-mail* | ***philippe.rohou@ercim.eu*** | |

** Contact details can only be changed by the Proposal Coordinator via the "Step 4 – Manage Your Related Parties" screen.*

EUROPEAN COMMISSION

7th Framework Programme for
Research, technological
Development and Demonstration

# A2.2 Participant #2

## Maaya

If your organisation has already registered for FP7, enter your Participant Identity Code ***959820433***

Legal Name ***Maaya***

Organisation short name ***Maaya***

## Administrative data (legal address)

Street name ***rue de Carouge*** Number ***104***

Town ***Genève 4***

Postal Code / Cedex ***1211***

Country ***CH***

Internet homepage ***http://www.maayajo.org/***

## Status of your organisation

Certain types of organisations benefit from special conditions under the FP7 participant rules. The Commission also collects data for statistical purposes.

The guidance notes will help you complete this section.

Please 'tick' the relevant box(es) if your organisation falls into one ore more of the following categories.

| | | |
|---|---|---|
| Non-profit organisation | ○ Yes | ◉ No |
| Public body | ◉ Yes | ○ No |
| Research organisation | ○ Yes | ◉ No |
| Higher or secondary education establishment | ○ Yes | ◉ No |

Main area of activity (NACE code) 92

EUROPEAN COMMISSION

7th Framework Programme for
Research, technological
Development and Demonstration

| | | |
|---|---|---|
| 1. Is your number of employees smaller than 250? (full time equivalent) | (•) Yes | ( ) No |
| 2. Is your annual turnover smaller than € 50 million? | (•) Yes | ( ) No |
| 3. Is your annual balance sheet total smaller than € 43 million? | (•) Yes | ( ) No |
| 4. Are you an autonomous legal entity? | (•) Yes | ( ) No |

You are NOT an SME if your answer to question 1 is "NO" and/or your answer to both questions 2 and 3 is "NO".

In all other cases, you might conform to the Commission's definition of an SME.

Please check the additional conditions given in the guidance notes to the forms.

Following this check, do you conform to the Commission's definition of an SME? ( ) Yes (•) No

# Dependencies with (an)other participant(s)

Are there dependencies between your organisation and (an)other participant(s) in this proposal? ( ) Yes (•) No

EUROPEAN COMMISSION

7th Framework Programme for
Research, technological
Development and Demonstration

# Contact point - Person in charge for participant #2

For the co-ordinator (Participant #1) this person is the one who the Commission will contact in the first instance.

Family name* ***Prado*** First name(s)* ***Daniel***

Title Mr.

(•) Male ( ) Female

Position in the organisation Executive Secretary

Department/Faculty/Institute/Laboratory name/... not applicable

**Address** ☐ Same as legal address

Street name 68ter, Ave Ledru-Rollin Number 68T

Town Le Perreux-sur-Marne Postal Code/Cedex 94170

Country FR

Phone1* + Phone2 + 33 0977195494

Fax + 33 0956721725 E-mail* ***dhprado@gmail.com***

** Contact details can only be changed by the Proposal Coordinator via the "Step 4 – Manage Your Related Parties" screen.*

EUROPEAN COMMISSION

7th Framework Programme for
Research, technological
Development and Demonstration

# A2.3 Participant #3

## UNIVERSITAT POLITECNICA DE CATALUNYA

| | |
|---|---|
| If your organisation has already registered for FP7, enter your Participant Identity Code | ***999976202*** |
| Legal Name | ***UNIVERSITAT POLITECNICA DE CATALUNYA*** |
| Organisation short name | ***UNIVERSITAT POLITECNICA DE CATALUNYA*** |

## Administrative data (legal address)

| | | | |
|---|---|---|---|
| Street name | ***Jordi Girona*** | Number | ***31*** |
| Town | ***BARCELONA*** | | |
| Postal Code / Cedex | ***08034*** | | |
| Country | ***ES*** | | |
| Internet homepage | ***www.upc.edu*** | | |

## Status of your organisation

Certain types of organisations benefit from special conditions under the FP7 participant rules. The Commission also collects data for statistical purposes.

The guidance notes will help you complete this section.

Please 'tick' the relevant box(es) if your organisation falls into one ore more of the following categories.

| | | |
|---|---|---|
| Non-profit organisation | (•) Yes | ( ) No |
| Public body | (•) Yes | ( ) No |
| Research organisation | (•) Yes | ( ) No |
| Higher or secondary education establishment | (•) Yes | ( ) No |

Main area of activity (NACE code): 80.3

EUROPEAN COMMISSION

7th Framework Programme for
Research, technological
Development and Demonstration

| | | |
|---|---|---|
| 1. Is your number of employees smaller than 250? (full time equivalent) | ○ Yes | ◉ No |
| 2. Is your annual turnover smaller than € 50 million? | ○ Yes | ◉ No |
| 3. Is your annual balance sheet total smaller than € 43 million? | ○ Yes | ◉ No |
| 4. Are you an autonomous legal entity? | ◉ Yes | ○ No |

You are NOT an SME if your answer to question 1 is "NO" and/or your answer to both questions 2 and 3 is "NO".

In all other cases, you might conform to the Commission's definition of an SME.

Please check the additional conditions given in the guidance notes to the forms.

Following this check, do you conform to the Commission's definition of an SME? ○ Yes ◉ No

# Dependencies with (an)other participant(s)

Are there dependencies between your organisation and (an)other participant(s) in this proposal? ○ Yes ◉ No

EUROPEAN COMMISSION

7th Framework Programme for
Research, technological
Development and Demonstration

# Contact point - Person in charge for participant #3

For the co-ordinator (Participant #1) this person is the one who the Commission will contact in the first instance.

Family name* ***Padro*** First name(s)* ***Lluis***

Title Dr.

(•) Male ( ) Female

Position in the organisation Associate Professor

Department/Faculty/Institute/Laboratory name/... TALP research center

**Address** ☐ Same as legal address

Street name Jordi Girona Number 1

Town Barcelona Postal Code/Cedex 08034

Country ES

Phone1* + Phone2 + 34 934134015

Fax + 34 934137833 E-mail* ***padro@lsi.upc.edu***

** Contact details can only be changed by the Proposal Coordinator via the "Step 4 – Manage Your Related Parties" screen.*

EUROPEAN COMMISSION

7th Framework Programme for
Research, technological
Development and Demonstration

# A2.4 Participant #4

## DIALOGIC INNOVATIE & INTERACTIE BV

| | |
|---|---|
| If your organisation has already registered for FP7, enter your Participant Identity Code | ***964264488*** |
| Legal Name | ***DIALOGIC INNOVATIE & INTERACTIE BV*** |
| Organisation short name | ***DIALOGIC INNOVATIE & INTERACTIE BV*** |

## Administrative data (legal address)

| | | | |
|---|---|---|---|
| Street name | ***HOOGHIEMSTRAPLEIN*** | Number | ***33-36*** |
| Town | ***UTRECHT*** | | |
| Postal Code / Cedex | ***3514AX*** | | |
| Country | ***NL*** | | |
| Internet homepage | ***http://www.dialogic.nl*** | | |

## Status of your organisation

Certain types of organisations benefit from special conditions under the FP7 participant rules. The Commission also collects data for statistical purposes.

The guidance notes will help you complete this section.

Please 'tick' the relevant box(es) if your organisation falls into one ore more of the following categories.

| | | |
|---|---|---|
| Non-profit organisation | ○ Yes | ◉ No |
| Public body | ○ Yes | ◉ No |
| Research organisation | ○ Yes | ◉ No |
| Higher or secondary education establishment | ○ Yes | ◉ No |

Main area of activity (NACE code): 74.1

EUROPEAN COMMISSION

7th Framework Programme for
Research, technological
Development and Demonstration

| | | |
|---|---|---|
| 1. Is your number of employees smaller than 250? (full time equivalent) | (•) Yes | ( ) No |
| 2. Is your annual turnover smaller than € 50 million? | (•) Yes | ( ) No |
| 3. Is your annual balance sheet total smaller than € 43 million? | (•) Yes | ( ) No |
| 4. Are you an autonomous legal entity? | (•) Yes | ( ) No |

You are NOT an SME if your answer to question 1 is "NO" and/or your answer to both questions 2 and 3 is "NO".

In all other cases, you might conform to the Commission's definition of an SME.

Please check the additional conditions given in the guidance notes to the forms.

Following this check, do you conform to the Commission's definition of an SME? (•) Yes ( ) No

# Dependencies with (an)other participant(s)

Are there dependencies between your organisation and (an)other participant(s) in this proposal? ( ) Yes (•) No

EUROPEAN COMMISSION

7th Framework Programme for
Research, technological
Development and Demonstration

# Contact point - Person in charge for participant #4

For the co-ordinator (Participant #1) this person is the one who the Commission will contact in the first instance.

Family name* ***te Velde*** First name(s)* ***Robbin***

Title Mr.

(•) Male ( ) Female

Position in the organisation Principal researcher

Department/Faculty/Institute/Laboratory name/... not applicable

**Address** ☐ Same as legal address

Street name Hooghiemstraplein Number 33-36

Town Utrecht Postal Code/Cedex 3514AX

Country NL

Phone1* + Phone2 + 31 302150580

Fax + 31 302150595 E-mail* ***tevelde@dialogic.nl***

** Contact details can only be changed by the Proposal Coordinator via the "Step 4 – Manage Your Related Parties" screen.*

EUROPEAN COMMISSION

7th Framework Programme for Research, technological Development and Demonstration

# A2.5 Participant #5

## CNRS

If your organisation has already registered for FP7, enter your Participant Identity Code ***999997930***

Legal Name ***CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE***

Organisation short name ***CNRS***

## Administrative data (legal address)

Street name ***Rue Michel -Ange*** Number ***3***

Town ***PARIS***

Postal Code / Cedex ***75794***

Country ***FR***

Internet homepage ***www.cnrs.fr***

## Status of your organisation

Certain types of organisations benefit from special conditions under the FP7 participant rules. The Commission also collects data for statistical purposes.

The guidance notes will help you complete this section.

Please 'tick' the relevant box(es) if your organisation falls into one ore more of the following categories.

| | | |
|---|---|---|
| Non-profit organisation | (•) Yes | ( ) No |
| Public body | (•) Yes | ( ) No |
| Research organisation | (•) Yes | ( ) No |
| Higher or secondary education establishment | ( ) Yes | (•) No |

Main area of activity (NACE code) 73.1

EUROPEAN COMMISSION

7th Framework Programme for
Research, technological
Development and Demonstration

| | | |
|---|---|---|
| 1. Is your number of employees smaller than 250? (full time equivalent) | ○ Yes | ◉ No |
| 2. Is your annual turnover smaller than € 50 million? | ○ Yes | ◉ No |
| 3. Is your annual balance sheet total smaller than € 43 million? | ○ Yes | ◉ No |
| 4. Are you an autonomous legal entity? | ◉ Yes | ○ No |

You are NOT an SME if your answer to question 1 is "NO" and/or your answer to both questions 2 and 3 is "NO".

In all other cases, you might conform to the Commission's definition of an SME.

Please check the additional conditions given in the guidance notes to the forms.

Following this check, do you conform to the Commission's definition of an SME? ○ Yes ◉ No

# Dependencies with (an)other participant(s)

Are there dependencies between your organisation and (an)other participant(s) in this proposal? ○ Yes ◉ No

EUROPEAN COMMISSION

7th Framework Programme for
Research, technological
Development and Demonstration

# Contact point - Person in charge for participant #5

For the co-ordinator (Participant #1) this person is the one who the Commission will contact in the first instance.

Family name* ***Lamel*** First name(s)* ***Lori***

Title Dr.

○ Male ◉ Female

Position in the organisation Senior research scientist

Department/Faculty/Institute/Laboratory name/... Laboratoire d'Informatique pour la Mechanique et les Sciences de l'Ingeni

**Address** ☐ Same as legal address

Street name Universite Paris Sud, bat 508 Number

Town Orsay Postal Code/Cedex 91403

Country FR

Phone1* + Phone2 + 33 169858183

Fax + 33 169858088 E-mail* ***lamel@limsi.fr***

** Contact details can only be changed by the Proposal Coordinator via the "Step 4 – Manage Your Related Parties" screen.*

EUROPEAN COMMISSION

7th Framework Programme for
Research, technological
Development and Demonstration

# A2.6 Participant #6

## KYOS

If your organisation has already registered for FP7, enter your Participant Identity Code ***990160093***

Legal Name ***KYOS SARL***

Organisation short name ***KYOS***

## Administrative data (legal address)

Street name ***avenue Rosemont*** Number ***12bis***

Town ***GENEVE***

Postal Code / Cedex ***1208***

Country ***CH***

Internet homepage ***http://www.kyos.ch***

## Status of your organisation

Certain types of organisations benefit from special conditions under the FP7 participant rules. The Commission also collects data for statistical purposes.

The guidance notes will help you complete this section.

Please 'tick' the relevant box(es) if your organisation falls into one ore more of the following categories.

| | | |
|---|---|---|
| Non-profit organisation | ○ Yes | ◉ No |
| Public body | ○ Yes | ◉ No |
| Research organisation | ○ Yes | ◉ No |
| Higher or secondary education establishment | ○ Yes | ◉ No |

Main area of activity (NACE code) 72

EUROPEAN COMMISSION

7th Framework Programme for
Research, technological
Development and Demonstration

| | | |
|---|---|---|
| 1. Is your number of employees smaller than 250? (full time equivalent) | (•) Yes | ( ) No |
| 2. Is your annual turnover smaller than € 50 million? | (•) Yes | ( ) No |
| 3. Is your annual balance sheet total smaller than € 43 million? | (•) Yes | ( ) No |
| 4. Are you an autonomous legal entity? | (•) Yes | ( ) No |

You are NOT an SME if your answer to question 1 is "NO" and/or your answer to both questions 2 and 3 is "NO".

In all other cases, you might conform to the Commission's definition of an SME.

Please check the additional conditions given in the guidance notes to the forms.

Following this check, do you conform to the Commission's definition of an SME? (•) Yes ( ) No

# Dependencies with (an)other participant(s)

Are there dependencies between your organisation and (an)other participant(s) in this proposal? ( ) Yes (•) No

EUROPEAN COMMISSION

7th Framework Programme for
Research, technological
Development and Demonstration

# Contact point - Person in charge for participant #6

For the co-ordinator (Participant #1) this person is the one who the Commission will contact in the first instance.

Family name* ***Jacquier*** First name(s)* ***Fabien***

Title Mr.

(•) Male ( ) Female

Position in the organisation Director

Department/Faculty/Institute/Laboratory name/... Geneva

**Address** ☒ Same as legal address

Street name avenue Rosemont Number 12bis

Town GENEVE Postal Code/Cedex 1208

Country CH

Phone1* + Phone2 + 41 786800425

Fax + E-mail* ***fabien.jacquier@kyos.ch***

** Contact details can only be changed by the Proposal Coordinator via the "Step 4 – Manage Your Related Parties" screen.*

EUROPEAN COMMISSION

7th Framework Programme for
Research, technological
Development and Demonstration

# A2.7 Participant #7

## Fraunhofer

If your organisation has already registered for FP7, enter your Participant Identity Code ***999984059***

Legal Name ***FRAUNHOFER-GESELLSCHAFT ZUR FOERDERUNG DER ANGEWANDTEN FORSCHU***

Organisation short name ***Fraunhofer***

## Administrative data (legal address)

Street name ***Hansastrasse*** Number ***27C***

Town ***MUNCHEN***

Postal Code / Cedex ***80686***

Country ***DE***

Internet homepage ***www.fraunhofer.de***

## Status of your organisation

Certain types of organisations benefit from special conditions under the FP7 participant rules. The Commission also collects data for statistical purposes.

The guidance notes will help you complete this section.

Please 'tick' the relevant box(es) if your organisation falls into one ore more of the following categories.

| | | |
|---|---|---|
| Non-profit organisation | (•) Yes | ( ) No |
| Public body | ( ) Yes | (•) No |
| Research organisation | (•) Yes | ( ) No |
| Higher or secondary education establishment | ( ) Yes | (•) No |

Main area of activity (NACE code) 73.1

EUROPEAN COMMISSION

7th Framework Programme for
Research, technological
Development and Demonstration

| | | |
|---|---|---|
| 1. Is your number of employees smaller than 250? (full time equivalent) | ○ Yes | ○ No |
| 2. Is your annual turnover smaller than € 50 million? | ○ Yes | ○ No |
| 3. Is your annual balance sheet total smaller than € 43 million? | ○ Yes | ○ No |
| 4. Are you an autonomous legal entity? | ○ Yes | ○ No |

You are NOT an SME if your answer to question 1 is "NO" and/or your answer to both questions 2 and 3 is "NO".

In all other cases, you might conform to the Commission's definition of an SME.

Please check the additional conditions given in the guidance notes to the forms.

Following this check, do you conform to the Commission's definition of an SME? ○ Yes ◉ No

# Dependencies with (an)other participant(s)

Are there dependencies between your organisation and (an)other participant(s) in this proposal? ○ Yes ◉ No

EUROPEAN COMMISSION

7th Framework Programme for
Research, technological
Development and Demonstration

# Contact point - Person in charge for participant #7

For the co-ordinator (Participant #1) this person is the one who the Commission will contact in the first instance.

Family name* ***Paass*** First name(s)* ***Gerhard***

Title Dr.

(•) Male ( ) Female

Position in the organisation Senior Scientist

Department/Faculty/Institute/Laboratory name/... Fraunhofer IAIS

**Address** ☐ Same as legal address

Street name Schloss Birlinghoven Number

Town Sankt Augustin Postal Code/Cedex 53754

Country DE

Phone1* + Phone2 + 49 2241 142689

Fax + 49 2241 1442689 E-mail* ***gerhard.paass@iais.fraunhofer.de***

** Contact details can only be changed by the Proposal Coordinator via the "Step 4 – Manage Your Related Parties" screen.*

EUROPEAN COMMISSION

7th Framework Programme for
Research, technological
Development and Demonstration

# A2.8 Participant #8

## CWI

If your organisation has already registered for FP7, enter your Participant Identity Code ***999653968***

Legal Name ***STICHTING CENTRUM VOOR WISKUNDE EN INFORMATICA***

Organisation short name ***CWI***

## Administrative data (legal address)

Street name ***Science Park*** Number ***123***

Town ***AMSTERDAM***

Postal Code / Cedex ***1098XG***

Country ***NL***

Internet homepage ***http://www.cwi.nl***

## Status of your organisation

Certain types of organisations benefit from special conditions under the FP7 participant rules. The Commission also collects data for statistical purposes.

The guidance notes will help you complete this section.

Please 'tick' the relevant box(es) if your organisation falls into one ore more of the following categories.

| | | |
|---|---|---|
| Non-profit organisation | (•) Yes | ( ) No |
| Public body | ( ) Yes | (•) No |
| Research organisation | (•) Yes | ( ) No |
| Higher or secondary education establishment | ( ) Yes | (•) No |

Main area of activity (NACE code) 73.1

EUROPEAN COMMISSION

7th Framework Programme for
Research, technological
Development and Demonstration

| | | |
|---|---|---|
| 1. Is your number of employees smaller than 250? (full time equivalent) | (•) Yes | ( ) No |
| 2. Is your annual turnover smaller than € 50 million? | (•) Yes | ( ) No |
| 3. Is your annual balance sheet total smaller than € 43 million? | (•) Yes | ( ) No |
| 4. Are you an autonomous legal entity? | (•) Yes | ( ) No |

You are NOT an SME if your answer to question 1 is "NO" and/or your answer to both questions 2 and 3 is "NO".

In all other cases, you might conform to the Commission's definition of an SME.

Please check the additional conditions given in the guidance notes to the forms.

Following this check, do you conform to the Commission's definition of an SME? ( ) Yes (•) No

# Dependencies with (an)other participant(s)

Are there dependencies between your organisation and (an)other participant(s) in this proposal? ( ) Yes (•) No

EUROPEAN COMMISSION

7th Framework Programme for
Research, technological
Development and Demonstration

# Contact point - Person in charge for participant #8

For the co-ordinator (Participant #1) this person is the one who the Commission will contact in the first instance.

Family name* ***De Vries*** First name(s)* ***Arjen***

Title Prof.

(•) Male ( ) Female

Position in the organisation Groupleader

Department/Faculty/Institute/Laboratory name/... Information Access (IA)

**Address** ☒ Same as legal address

Street name Science Park Number 123

Town AMSTERDAM Postal Code/Cedex 1098XG

Country NL

Phone1* + Phone2 + 31 20 592 4306

Fax + E-mail* ***arjen@acm.org***

** Contact details can only be changed by the Proposal Coordinator via the "Step 4 – Manage Your Related Parties" screen.*

EUROPEAN COMMISSION

7th Framework Programme for Research, technological Development and Demonstration

# A2.9 Participant #9

## Vocapia Research

If your organisation has already registered for FP7, enter your Participant Identity Code ***951433037***

Legal Name ***VOCAPIA RESEARCH***

Organisation short name ***Vocapia Research***

## Administrative data (legal address)

Street name ***rue Jean Rostand*** Number ***28***

Town ***Orsay***

Postal Code / Cedex ***91400***

Country ***FR***

Internet homepage ***www.vocapia.com***

## Status of your organisation

Certain types of organisations benefit from special conditions under the FP7 participant rules. The Commission also collects data for statistical purposes.

The guidance notes will help you complete this section.

Please 'tick' the relevant box(es) if your organisation falls into one ore more of the following categories.

| | | |
|---|---|---|
| Non-profit organisation | ○ Yes | ◉ No |
| Public body | ○ Yes | ◉ No |
| Research organisation | ○ Yes | ◉ No |
| Higher or secondary education establishment | ○ Yes | ◉ No |

Main area of activity (NACE code) 72

EUROPEAN COMMISSION

7th Framework Programme for
Research, technological
Development and Demonstration

| | | |
|---|---|---|
| 1. Is your number of employees smaller than 250? (full time equivalent) | (•) Yes | ( ) No |
| 2. Is your annual turnover smaller than € 50 million? | (•) Yes | ( ) No |
| 3. Is your annual balance sheet total smaller than € 43 million? | (•) Yes | ( ) No |
| 4. Are you an autonomous legal entity? | (•) Yes | ( ) No |

You are NOT an SME if your answer to question 1 is "NO" and/or your answer to both questions 2 and 3 is "NO".

In all other cases, you might conform to the Commission's definition of an SME.

Please check the additional conditions given in the guidance notes to the forms.

Following this check, do you conform to the Commission's definition of an SME? (•) Yes ( ) No

# Dependencies with (an)other participant(s)

Are there dependencies between your organisation and (an)other participant(s) in this proposal? ( ) Yes (•) No

EUROPEAN COMMISSION

7th Framework Programme for
Research, technological
Development and Demonstration

# Contact point - Person in charge for participant #9

For the co-ordinator (Participant #1) this person is the one who the Commission will contact in the first instance.

Family name* ***Vieru*** First name(s)* ***Bianca***

Title Dr.

○ Male ◉ Female

Position in the organisation R&D engineer

Department/Faculty/Institute/Laboratory name/... not applicable

**Address** ☒ Same as legal address

Street name rue Jean Rostand Number 28

Town Orsay Postal Code/Cedex 91400

Country FR

Phone1* + Phone2 + 33 184169617

Fax + 33 160195494 E-mail* ***vieru@vocapia.com***

** Contact details can only be changed by the Proposal Coordinator via the "Step 4 – Manage Your Related Parties" screen.*

EUROPEAN COMMISSION

7th Framework Programme for Research, technological Development and Demonstration

# A2.10 Participant #10

## Nielsen

If your organisation has already registered for FP7, enter your Participant Identity Code ***954782059***

Legal Name ***Nielsen Media Research GmbH***

Organisation short name ***Nielsen***

## Administrative data (legal address)

Street name ***Sachsenstrasse 16*** Number

Town ***Hamburg***

Postal Code / Cedex ***20097***

Country ***DE***

Internet homepage ***www.nielsen.com***

## Status of your organisation

Certain types of organisations benefit from special conditions under the FP7 participant rules. The Commission also collects data for statistical purposes.

The guidance notes will help you complete this section.

Please 'tick' the relevant box(es) if your organisation falls into one ore more of the following categories.

| | | |
|---|---|---|
| Non-profit organisation | ○ Yes | ◉ No |
| Public body | ○ Yes | ◉ No |
| Research organisation | ○ Yes | ◉ No |
| Higher or secondary education establishment | ○ Yes | ◉ No |

Main area of activity (NACE code) 93

EUROPEAN COMMISSION

7th Framework Programme for
Research, technological
Development and Demonstration

| | | |
|---|---|---|
| 1. Is your number of employees smaller than 250? (full time equivalent) | ○ Yes | ◉ No |
| 2. Is your annual turnover smaller than € 50 million? | ○ Yes | ◉ No |
| 3. Is your annual balance sheet total smaller than € 43 million? | ○ Yes | ◉ No |
| 4. Are you an autonomous legal entity? | ◉ Yes | ○ No |

You are NOT an SME if your answer to question 1 is "NO" and/or your answer to both questions 2 and 3 is "NO".

In all other cases, you might conform to the Commission's definition of an SME.

Please check the additional conditions given in the guidance notes to the forms.

Following this check, do you conform to the Commission's definition of an SME? ○ Yes ◉ No

# Dependencies with (an)other participant(s)

Are there dependencies between your organisation and (an)other participant(s) in this proposal? ○ Yes ◉ No

EUROPEAN COMMISSION

7th Framework Programme for
Research, technological
Development and Demonstration

# Contact point - Person in charge for participant #10

For the co-ordinator (Participant #1) this person is the one who the Commission will contact in the first instance.

Family name* ***Lamsfuss*** First name(s)* ***René***

Title Mr.

(•) Male ( ) Female

Position in the organisation Commercial Director

Department/Faculty/Institute/Laboratory name/... German Nielsen Office

**Address** ☒ Same as legal address

Street name Sachsenstrasse 16 Number

Town Hamburg Postal Code/Cedex 20097

Country DE

Phone1* + Phone2 + 49 1622007553

Fax + E-mail* ***rene.lamsfuss@nielsen.com***

** Contact details can only be changed by the Proposal Coordinator via the "Step 4 – Manage Your Related Parties" screen.*

EUROPEAN COMMISSION

7th Framework Programme for
Research, technological
Development and Demonstration

# A3.1.1 Budget #1

## ERCIM

In FP7, there are different methods for calculating indirect costs. The various options are explained in the guidance notes. Please be aware that not all options are available to all types of organisations.

Method: Actual indirect costs

My legal entity is established in an ICPC and I shall use the lump sum funding method
(If yes, please fill below the lump sum row only. If no, please do not use the lump sum row)

○ Yes ◉ No

Type of Activity

| | RTD<br>up to 50 or 75% * | Demonstration<br>up to 50% | Management<br>up to 100% | Total |
|---|---|---|---|---|
| Personnel costs (in €) | | | 125 465 € | 125 465 € |
| Subcontracting (in €) | | | | |
| Other direct costs (in €) | | | 27 200 € | 27 200 € |
| Indirect costs (in €) | | | 132 570 € | 132 570 € |
| Lump sum, flat-rate or scale of unit (option only for ICPC) (in €) | | | | |
| Total budget (in €) | | | 285 235 € | 285 235 € |
| Requested EC contribution (in €) | | | 285 235 € | 285 235 € |
| **Total Receipts (in €)** | | | | 0 € |

EUROPEAN COMMISSION

7th Framework Programme for
Research, technological
Development and Demonstration

# A3.1.2 Budget #2

## Maaya

In FP7, there are different methods for calculating indirect costs. The various options are explained in the guidance notes. Please be aware that not all options are available to all types of organisations.

Method: Actual indirect costs using a simplified method

My legal entity is established in an ICPC and I shall use the lump sum funding method
(If yes, please fill below the lump sum row only. If no, please do not use the lump sum row)

○ Yes ◉ No

Type of Activity

| | RTD<br>up to 50 or 75% * | Demonstration<br>up to 50% | Management<br>up to 100% | Total |
|---|---|---|---|---|
| Personnel costs (in €) | 249 750 € | | 83 250 € | 333 000 € |
| Subcontracting (in €) | | | | |
| Other direct costs (in €) | 10 800 € | | | 10 800 € |
| Indirect costs (in €) | 130 275 € | 0 € | 41 625 € | 171 900 € |
| Lump sum, flat-rate or scale of unit (option only for ICPC) (in €) | | | | |
| Total budget (in €) | 390 825 € | 0 € | 124 875 € | 515 700 € |
| Requested EC contribution (in €) | 293 118 € | 0 € | 124 875 € | 417 993 € |
| **Total Receipts (in €)** | | | | 0 € |

EUROPEAN COMMISSION

7th Framework Programme for
Research, technological
Development and Demonstration

# A3.1.3 Budget #3

## UNIVERSITAT POLITECNICA DE CATALUNYA

In FP7, there are different methods for calculating indirect costs. The various options are explained in the guidance notes. Please be aware that not all options are available to all types of organisations.

Method: Actual indirect costs

My legal entity is established in an ICPC and I shall use the lump sum funding method
(If yes, please fill below the lump sum row only. If no, please do not use the lump sum row)

○ Yes ◉ No

Type of Activity

| | RTD<br>up to 50 or 75% * | Demonstration<br>up to 50% | Management<br>up to 100% | Total |
|---|---|---|---|---|
| Personnel costs (in €) | 104 500 € | | 1 425 € | 105 925 € |
| Subcontracting (in €) | | | | |
| Other direct costs (in €) | 10 800 € | | | 10 800 € |
| Indirect costs (in €) | 79 953 € | | 1 090 € | 81 043 € |
| Lump sum, flat-rate or scale of unit (option only for ICPC) (in €) | | | | |
| Total budget (in €) | 195 253 € | | 2 515 € | 197 768 € |
| Requested EC contribution (in €) | 146 439 € | 0 € | 2 515 € | 148 954 € |
| **Total Receipts (in €)** | | | | 0 € |

EUROPEAN COMMISSION

7th Framework Programme for
Research, technological
Development and Demonstration

# A3.1.4 Budget #4

## DIALOGIC INNOVATIE & INTERACTIE BV

In FP7, there are different methods for calculating indirect costs. The various options are explained in the guidance notes. Please be aware that not all options are available to all types of organisations.

Method: Specific flat rate 60%

My legal entity is established in an ICPC and I shall use the lump sum funding method
(If yes, please fill below the lump sum row only. If no, please do not use the lump sum row)

○ Yes ◉ No

Type of Activity

| | RTD<br>up to 50 or 75% * | Demonstration<br>up to 50% | Management<br>up to 100% | Total |
|---|---|---|---|---|
| Personnel costs (in €) | 298 294 € | | 3 086 € | 301 380 € |
| Subcontracting (in €) | | | 2 000 € | 2 000 € |
| Other direct costs (in €) | 10 800 € | | | 10 800 € |
| Indirect costs (in €) | 185 456 € | 0 € | 1 851 € | 187 307 € |
| Lump sum, flat-rate or scale of unit (option only for ICPC) (in €) | | | | |
| Total budget (in €) | 494 550 € | 0 € | 6 937 € | 501 487 € |
| Requested EC contribution (in €) | 370 912 € | | 6 937 € | 377 849 € |
| **Total Receipts (in €)** | | | | 0 € |

EUROPEAN COMMISSION

7th Framework Programme for
Research, technological
Development and Demonstration

# A3.1.5 Budget #5

## CNRS

In FP7, there are different methods for calculating indirect costs. The various options are explained in the guidance notes. Please be aware that not all options are available to all types of organisations.

Method: Specific flat rate 60%

My legal entity is established in an ICPC and I shall use the lump sum funding method
(If yes, please fill below the lump sum row only. If no, please do not use the lump sum row)

○ Yes ◉ No

Type of Activity

| | RTD<br>up to 50 or 75% * | Demonstration<br>up to 50% | Management<br>up to 100% | Total |
|---|---|---|---|---|
| Personnel costs (in €) | 175 500 € | | 1 950 € | 177 450 € |
| Subcontracting (in €) | | | | |
| Other direct costs (in €) | 10 800 € | | | 10 800 € |
| Indirect costs (in €) | 111 780 € | 0 € | 1 170 € | 112 950 € |
| Lump sum, flat-rate or scale of unit (option only for ICPC) (in €) | | | | |
| Total budget (in €) | 298 080 € | 0 € | 3 120 € | 301 200 € |
| Requested EC contribution (in €) | 223 560 € | | 3 120 € | 226 680 € |
| **Total Receipts (in €)** | | | | 0 € |

EUROPEAN COMMISSION

7th Framework Programme for
Research, technological
Development and Demonstration

# A3.1.6 Budget #6

## KYOS

In FP7, there are different methods for calculating indirect costs. The various options are explained in the guidance notes. Please be aware that not all options are available to all types of organisations.

Method: Specific flat rate 60%

My legal entity is established in an ICPC and I shall use the lump sum funding method
(If yes, please fill below the lump sum row only. If no, please do not use the lump sum row)

○ Yes ◉ No

Type of Activity

| | RTD<br>up to 50 or 75% * | Demonstration<br>up to 50% | Management<br>up to 100% | Total |
|---|---|---|---|---|
| Personnel costs (in €) | 96 000 € | | 2 400 € | 98 400 € |
| Subcontracting (in €) | | | | |
| Other direct costs (in €) | 10 800 € | | | 10 800 € |
| Indirect costs (in €) | 64 080 € | 0 € | 1 440 € | 65 520 € |
| Lump sum, flat-rate or scale of unit (option only for ICPC) (in €) | | | | |
| Total budget (in €) | 170 880 € | 0 € | 3 840 € | 174 720 € |
| Requested EC contribution (in €) | 128 160 € | | 3 840 € | 132 000 € |
| **Total Receipts (in €)** | | | | 0 € |

EUROPEAN COMMISSION

7th Framework Programme for
Research, technological
Development and Demonstration

# A3.1.7 Budget #7

## Fraunhofer

In FP7, there are different methods for calculating indirect costs. The various options are explained in the guidance notes. Please be aware that not all options are available to all types of organisations.

Method: Actual indirect costs

My legal entity is established in an ICPC and I shall use the lump sum funding method
(If yes, please fill below the lump sum row only. If no, please do not use the lump sum row)

○ Yes ◉ No

Type of Activity

| | RTD<br>up to 50 or 75% * | Demonstration<br>up to 50% | Management<br>up to 100% | Total |
|---|---|---|---|---|
| Personnel costs (in €) | 527 655 € | | 1 820 € | 529 475 € |
| Subcontracting (in €) | | | 4 000 € | 4 000 € |
| Other direct costs (in €) | 50 800 € | | | 50 800 € |
| Indirect costs (in €) | 405 323 € | | 1 275 € | 406 598 € |
| Lump sum, flat-rate or scale of unit (option only for ICPC) (in €) | | | | |
| Total budget (in €) | 983 778 € | | 7 095 € | 990 873 € |
| Requested EC contribution (in €) | 737 833 € | 0 € | 7 095 € | 744 928 € |
| **Total Receipts (in €)** | | | | 0 € |

EUROPEAN COMMISSION

7th Framework Programme for
Research, technological
Development and Demonstration

# A3.1.8 Budget #8

## CWI

In FP7, there are different methods for calculating indirect costs. The various options are explained in the guidance notes. Please be aware that not all options are available to all types of organisations.

Method: Actual indirect costs

My legal entity is established in an ICPC and I shall use the lump sum funding method
(If yes, please fill below the lump sum row only. If no, please do not use the lump sum row)

○ Yes ◉ No

Type of Activity

| | RTD<br>up to 50 or 75% * | Demonstration<br>up to 50% | Management<br>up to 100% | Total |
|---|---|---|---|---|
| Personnel costs (in €) | 112 392 € | | 1 606 € | 113 998 € |
| Subcontracting (in €) | | | | |
| Other direct costs (in €) | 10 800 € | | | 10 800 € |
| Indirect costs (in €) | 69 237 € | | 989 € | 70 226 € |
| Lump sum, flat-rate or scale of unit (option only for ICPC) (in €) | | | | |
| Total budget (in €) | 192 429 € | | 2 595 € | 195 024 € |
| Requested EC contribution (in €) | 144 321 € | | 2 595 € | 146 916 € |
| **Total Receipts (in €)** | | | | 0 € |

EUROPEAN COMMISSION

7th Framework Programme for
Research, technological
Development and Demonstration

# A3.1.9 Budget #9

## Vocapia Research

In FP7, there are different methods for calculating indirect costs. The various options are explained in the guidance notes. Please be aware that not all options are available to all types of organisations.

Method: Specific flat rate 60%

My legal entity is established in an ICPC and I shall use the lump sum funding method
(If yes, please fill below the lump sum row only. If no, please do not use the lump sum row)

○ Yes ◉ No

Type of Activity

| | RTD<br>up to 50 or 75% * | Demonstration<br>up to 50% | Management<br>up to 100% | Total |
|---|---|---|---|---|
| Personnel costs (in €) | 201 000 € | | 2 010 € | 203 010 € |
| Subcontracting (in €) | | | 2 000 € | 2 000 € |
| Other direct costs (in €) | 10 800 € | | | 10 800 € |
| Indirect costs (in €) | 127 080 € | | 1 206 € | 128 286 € |
| Lump sum, flat-rate or scale of unit (option only for ICPC) (in €) | | | | |
| Total budget (in €) | 338 880 € | | 5 216 € | 344 096 € |
| Requested EC contribution (in €) | 254 160 € | | 5 216 € | 259 376 € |
| **Total Receipts (in €)** | | | | 0 € |

EUROPEAN COMMISSION

7th Framework Programme for
Research, technological
Development and Demonstration

# A3.1.10 Budget #10

## Nielsen

In FP7, there are different methods for calculating indirect costs. The various options are explained in the guidance notes. Please be aware that not all options are available to all types of organisations.

Method: Actual indirect costs

My legal entity is established in an ICPC and I shall use the lump sum funding method
(If yes, please fill below the lump sum row only. If no, please do not use the lump sum row)

○ Yes ◉ No

Type of Activity

| | RTD<br>up to 50 or 75% * | Demonstration<br>up to 50% | Management<br>up to 100% | Total |
|---|---|---|---|---|
| Personnel costs (in €) | 320 450 € | | 33 150 € | 353 600 € |
| Subcontracting (in €) | | | | |
| Other direct costs (in €) | 10 800 € | | | 10 800 € |
| Indirect costs (in €) | 99 375 € | | 9 945 € | 109 320 € |
| Lump sum, flat-rate or scale of unit (option only for ICPC) (in €) | | | | |
| Total budget (in €) | 430 625 € | | 43 095 € | 473 720 € |
| Requested EC contribution (in €) | 215 312 € | | 43 095 € | 258 407 € |
| **Total Receipts (in €)** | | | | 0 € |

EUROPEAN COMMISSION

7th Framework Programme for
Research, technological
Development and Demonstration

# A3.2: Budget

## Estimated budget in EUR (whole of the project)

| Nr. | Organisation Short Name | Organisation country | RTD | Demonstration | Management | Total | Total receipts | Requested EU contributions |
|---|---|---|---|---|---|---|---|---|
| 1 | ERCIM | FR | | | 285 235 | 285 235 | 0 | 285 235 |
| 2 | Maaya | CH | 390 825 | 0 | 124 875 | 515 700 | 0 | 417 993 |
| 3 | UNIVERSITAT POLITECNICA DE CATALUNYA | ES | 195 253 | | 2 515 | 197 768 | 0 | 148 954 |
| 4 | DIALOGIC INNOVATIE & INTERACTIE BV | NL | 494 550 | 0 | 6 937 | 501 487 | 0 | 377 849 |
| 5 | CNRS | FR | 298 080 | 0 | 3 120 | 301 200 | 0 | 226 680 |
| 6 | KYOS | CH | 170 880 | 0 | 3 840 | 174 720 | 0 | 132 000 |
| 7 | Fraunhofer | DE | 983 778 | | 7 095 | 990 873 | 0 | 744 928 |
| 8 | CWI | NL | 192 429 | | 2 595 | 195 024 | 0 | 146 916 |
| 9 | Vocapia Research | FR | 338 880 | | 5 216 | 344 096 | 0 | 259 376 |
| 10 | Nielsen | DE | 430 625 | | 43 095 | 473 720 | 0 | 258 407 |

EUROPEAN COMMISSION

7th Framework Programme for
Research, technological
Development and Demonstration

| | | | | | | |
|---|---|---|---|---|---|---|
| Total | 3 495 300 | 0 | 484 523 | 3 979 823 | 0 | 2 998 338 |

**Small or medium scale focused project**

**(STREP)**

# SEMACORE

**SE**mantic **M**ultimedia **A**nalysis Performed **C**r**O**ss-Language in Near **RE**altime

# PART B

**Objective ICT-2013.4.1 Content analytics and language technologies**
**Target outcome a) Cross-media content analytics.**

Keywords: multimedia, sentiment analysis, language diversity, cross-language information extraction, multilingualism, content extraction from multimedia documents, big data analytics, extendable analysis workflow

Date of preparation: 15/01/2013 Version number: V25

| Beneficiary no. | Beneficiary name | Short name | Country |
|---|---|---|---|
| 1 (coordinator) | The European Research Consortium for Informatics and Mathematics | ERCIM | FR |
| 2 | World Network for Linguistic Diversity | MAAYA | CH |
| 3 | Universitat Politecnica de Catalunya | UPC | ES |
| 4 | Dialogic Innovation & Interaction | DIALOGIC | NL |
| 5 | Centre National de la Recherche Scientifique | CNRS | FR |
| 6 | Kyos IT Security | KYOS | CH |
| 7 | Fraunhofer-Gesellschaft zur Förderung der angewandten Forschung e.V. | FRAUNHOFER | DE |
| 8 | Dutch National Centre of Mathematics and Computer Science | CWI | NL |
| 9 | Vocapia Research | VOCAPIA | FR |
| 10 | Nielsen Media Research GmbH | NIELSEN | DE |

Coordinator's name: Philippe Rohou
Coordinator's e-mail: philippe.rohou@ercim.eu
Coordinator's fax: +33 (0)4 92 38 78 22

## Proposal abstract

The widespread adoption of smartphones and social media and the availability of Big Data processing capabilities for the first time facilitate the comprehensive analysis of citizen's statements and communication. In particular a multilingual evaluation of the contents of web documents has become possible, offering huge potential societal and economic benefits.

SEMACORE will advance the semantic characterization of the web in three crucial aspects: (1) develop methods to analyse content and opinions in multiple languages with uniform semantic definitions, (2) simultaneous analysis of contents and opinions in text and speech from audio/video (3) near real-time exploration of contents in multiple languages and media to cover current events.

The goal of the SEMACORE project is to radically advance the analysis of cross-lingual contents and opinions by developing innovative technologies, methodologies and systems that will put new capabilities in the hand of political and economic decision makers. This will offer a joint and dynamic perception of the opinions of European citizens.

SEMACORE brings together a strong group of researchers with domain experts in three representative use cases on the joint analysis of language, content, and opinions in web documents. The World Network for Linguistic Diversity MAAYA will analyse language use and contents in small countries, international organizations and digital libraries. Nielsen, a leading global media and market research company, will evaluate contents and opinions for an international organization and a company. The technology on speech recognition, machine translation and multilingual content/opinion mining will be provided and developed further by CNRS, UPC, Fraunhofer, and CWI. DIALOGIC will monitor a representative volunteer panel of PC and smartphone users. SEMACORE will provide a standardized workflow for including new languages in the joint analysis opening outstanding avenues for exploitation in business and policy.

# Table of contents

# Figures and Tables

# B 1. Concept and objectives, progress beyond state-of-the-art, S/T methodology and work plan

## B 1.1 Concept and project objective(s)

**A communication revolution.** Today's technological advances are fuelling a virtual explosion on the quantity, quality, and variety of information that is becoming available. The mega trends that drive this proposal are three recent revolutionary technologies:

1. **The widespread adoption of smartphones**: With the proliferation of smartphones, tablets, and other mobile computing devices, we enter an era where people can share instantaneously aspects of their lives, creating virtual communities with direct connections to the real world, and becoming both data producers and data consumers at the same time. In the first quarter of 2012 47.6% of people in the EU5 (France, Germany, Italy, Spain, UK) used a smartphone.
2. **The extent of social networks:** Social networks, such as Twitter or Facebook, have significantly changed the way humans interact. Especially young users are always online and exchange short messages, photos, and comments with their virtual communities. In the EU 80% among those aged 16-24 use the internet for posting messages to chat sites, blogs and social networks.[1] These networks are also used to disseminate news, entertainment, opinions on issues and products, and even to organize social action.
3. **Big Data analysis**: Huge volumes of data generated by traditional business activities and from new sources such as smartphones and social media are collected in data warehouses and the cloud. This structured and unstructured data is analysed looking for hidden patterns, trends or other insights that can be used to better tailor products and services to customers, anticipate demand or improve performance.

**The challenge of language diversity.** Although the European Union exists for more than 50 years the perception and discussion of political and economic issues usually happens in national media and discussion forums using the 23 officially recognized languages and more than 60 indigenous regional and minority languages. The new revolutionary technologies now facilitate the establishment of a European public, and a unified view on the discussion of political, economic and cultural issues in Europe across language barriers. This can be achieved without reducing European multilingualism, which is a key ingredient of cultural identity. This truly visionary application is achieved by building on these three technologies.

### B 1.1.1 The SEMACORE vision

**The vision of SEMACORE is to develop technologies for the fundamental advances in communicating information across language barriers. We will develop technologies for the cross-language extraction of facts and opinions from the web and social networks for their aggregation, integration and instantaneous publication to interested stakeholders, communities and customers.**

**End-user driven scenarios for multilingual semantic analysis of web contents.** Our efforts will be realized in the context of carefully selected use cases that aim to build fundamental capabilities in evaluating multilingual web content across different media. The use cases chosen, (i) employ a broad spectrum of technology requirements, (ii) address important end-user driven problems, and (iii) have the potential for enormous impact on real-world problems; thus ensuring broad applicability of results.

[1] http://www.newmediatrendwatch.com/regional-overview/103-europe?showall=1

Reliable, fast and effective analysis of the language, the attitudes and the opinions of citizens and consumers is essential for every public institution and every company in the market. The use cases we consider benefit greatly from the introduction of information technology and machine learning methods that allow the full utilization of all data. However, current solutions are severely limited in their ability to take advantage from the huge amounts of data available in the web and especially in social networks. Main media monitoring services collect their data electronically but often still annotate contents manually. Moreover there is usually no uniform cross-language and cross-media analysis of user statements. The reason is the lack of techniques that can both deal with the volume of available data, and the limited performance of automatic approaches up to now. Our proposal aims at carrying out groundbreaking research to deal with heterogeneous sources of information in a unified framework and at the same time accomplish a high level of reliability in semantic interpretation. Next we provide a short description of each use case that will be considered. Extended descriptions with details of the associated data are given in section B1.3.4 under paragraph WP7.

**Measuring actual language use in the web in different regions and contexts.** Within the EU there are more than 83 official and indigenous regional and minority languages and in addition many languages spoken by immigrants. The EU is committed to safeguarding this linguistic diversity for reasons of cultural identity and social integration and cohesion. In this use case SEMACORE will measure the use of different languages in the web. In contrast to other surveys[2] we will not only measure the language use but also the content communicated in different languages and media with respect to genre and content categories. In addition the extent of opinionated and factual information will be determined. Because of budget limitations these surveys will only be performed for small countries (Ireland, Luxembourg, Malta, Netherlands). There are estimates that 90% of the web is not covered by major search engines. SEMACORE will identify which language groups, national sites, media, genre and content categories are under-represented in these indexes. SEMACORE will also evaluate if for "small" languages specific relevant content is not existent or accessible inducing a "linguistic divide" such that citizens only speaking these languages have less chances to take advantage of the economic, educational and professional opportunities created by an integrated Europe. These results on the one hand may be used by the EC and national governments to provide additional content as well search/translation infrastructure in these languages. On the other hand industry may realize market opportunities by promoting goods and services in "small" languages which are currently neglected. Native speakers of these languages often had difficulty finding language technologies for their needs.

**Multilingual analysis of web content and opinions for an international public agency.** Because of the massive use of the internet in daily life it is vital for public stakeholders to promote their activities on the web and monitor the mentions of public organisations, political issues, or politicians in web documents and social networks. Existing tools, however, do not cover multiple languages and media. In this use case SEMACORE will measure the buzz around a keyword of interest with uniform semantics across different languages for a public agency. As a baseline we will evaluate the correlation of the disambiguated keyword in different languages with other features like topic distribution, public content and genre categories, time and the presence of opinion words. Even if the quality of translation and speech recognition is imperfect, this approach promises acceptable results. A more advanced multilingual opinion mining will extract the semantic relation between keywords and opinions offering more focussed results than correlational analysis but also requiring training data of higher quality. The detailed feedback provided by these analyses will allow the agency to adapt its communication strategy and adjust its activities.

**Multilingual analysis of web content and opinions for an international company.** About 47% of the Europeans will consult social media websites or online product reviews before making purchases of

[2] European Commission: Special Eurobarometer 386 – Europeans and their Languages. June 2012.

entertainment, home electronics, travel & leisure or appliances in 2013[3]. Hence the analysis of the reaction of customers to products and brands in the web is a decisive factor for the economic success of a company. In this use case SEMACORE will apply multilingual content analysis and opinion mining for an international manufacturer interested in the sentiment of web users towards its brands and products. As in the previous use case a baseline correlational approach as well as opinion mining will be developed. However, because of the different domains the training data for parts of the content analysis will be different. The analysis will provide information on the temporal and regional trends in the spread of brands and products and its association with content and genre categories. The opinion mining is designed to process high volumes of unstructured consumer generated stories and will elicit the particular positive and negative attributes of a product defined with uniform semantics for different languages and media. The company will be able to assess customer satisfaction and problem categories with a specific product, appreciate the effect of a recent advertising campaign on the product perception and evaluate the differences in product image between different countries, languages regions and web genres. This use case will be executed by Nielsen whose reference list for Social Media analysis is including for example Procter & Gamble, Disney, Honda, Western Union, 3M, and BBVA.

The goal of SEMACORE is a significant improvement in the utility of automated systems to analyse multilingual and multimodal web contents and put new capabilities in the hands of public agencies active in language politics or political issues as well as private companies promoting their products. Given the increasing amount of complexity of data and the diversity of different languages and media the challenge is very large. To achieve a major improvement, we have to integrate a large Big Data infrastructure with advanced speech recognition, machine translation and sophisticated semantic multilingual machine learning approaches still to be developed. Our goal is ambitious and the stakes are high. However such an approach is necessary to truly exploit the opportunities offered by the availability of web and social networking data together with new advances in speech recognition, machine translation and machine learning. Considering these factors simultaneously has not been done in any depth, yet the interplay has significant implications. Finally, we will build a real system. We plan to create an infrastructure that will provide a long-term support for building, maintaining, and improving such systems and provide a standardized workflow to include more languages.

**The need for innovation.** To realize this goal we have to innovate on several fronts, and develop novel techniques to handle massive data, novel techniques to design machine translation and speech recognition for non-official European Union languages and especially to integrate machine translation, latent variable models and information extraction in a novel way. The proposed partners bring state of the art expertise in distributed stream algorithms, distributed data mining, machine learning techniques, natural language processing, machine translation, speech recognition, real-time and fault-tolerant distributed systems, distributed programming, MapReduce, media analysis and market research.

### B 1.1.2 Project Objectives

The SEMACORE project aims at creating an open, powerful and reusable solution and developing the capabilities for fundamental advances in real-time cross-lingual fact and opinion mining and aggregation. To achieve our goal the project has the following innovation objectives:

**Objective 1: Revolutionize the measurement of language and opinions on the web by close to real-time monitoring** using massive streaming of heterogeneous and multimodal data. Language use, content categories as well as opinions are automatically extracted for different media like web pages, news, blogs, social networks, podcasts, YouTube, and the personal communication of volunteers. The analysis covers the six most important European languages (Dutch, English, French, German, Italian,

[3] Source: Nielsen. State of the media: The Social Media Report 2012. http://www.nielsen.com/us/en/insights/reports-downloads.html

Spanish) as well as one minority (Catalan) and two immigrant languages (Arabic, Turkish). See the language matrix on Figure 1 for details. The methodology ensures that the semantics of annotations are identical over the different languages and yields better results than the investigation of single languages. It allows the instantaneous analysis of the reaction of citizens on ongoing events like election campaigns, international championships and product launches. WP 7 describes the use cases that will realize Objective 1.

**Objective 2: To develop novel efficient and reusable methodologies for the detection of language, content and opinions on the web across multiple languages in text, websites, speech, video, and personal communication.** Audio data from speech and video will be transcribed to text by advanced speaker-independent speech recognizers. As a baseline technology state-of-the-art statistical machine translation will be used to project training resources to other languages where subsequently state of the art classifiers and opinion models are trained. In addition cross-lingual latent variable models will be developed automatically capturing the semantics of words by cross-language concepts connected to Wikipedia entries. As a radically new technology for grasping the relation between aspects and opinions we will develop multilingual deep learning models representing the sequence and structure of meaning annotations. The algorithms will be reusable and can easily be adapted to new languages. The work described in WP5 and WP6 will realize Objective 2.

**Objective 3: To develop an adaptive, scalable and dependable, real-time infrastructure for multimodal and multilingual opinion mining** facilitating a fast adoption of the proposed methodology we focus on building a scalable system that simplifies the process of crawling, extraction, aggregation and analysis. We develop a real-time distributed programming framework that offers MapReduce functionality, provides end-to-end near real-time and reliable delivery of continuous streams of data, while hiding the underlying difficult issues of distributed computing including concurrency, data distribution and real-time delivery from the end-user. The storage and processing system combines a streaming engine for fast online analysis of data streams with a big data warehouse for analysing and processing historic data and annotations. Our objective is to provide easy-to-use and practical software modules that enable efficient communication of the different SEMACORE components, fast processing of streaming data and powerful in-depth analysis of data within the warehouse. The work described in WP3 will realize Objective 3.

A major objective of SEMACORE is to support the adoption and use of the technologies that we will develop:

**Objective 4:** To ensure reusability and facilitate faster adoption and extension of the proposed methodology, we will provide a **standardized workflow for including new languages in the joint analysis**. This will include recommendations for the training of new speech recognizers and machine translators and will be based on the reuse of existing training data for content analysis. To facilitate faster adoption our second focus will be the development of a reference architecture and suitable APIs, so that different user communities and types of applications from the original use cases can apply the SEMACORE technology in their business. Objective 4 will be realized by the work in WP3, WP5, WP6, and WP7.

### B 1.1.3 Project outcome

The outcome of the SEMACORE project will be:

- an integrated Big Data infrastructure offering real-time and batch processing functionality,
- a comprehensive data storage for annotated documents and analysis results,
- characterization of the actual language use in the web in restricted domains,
- a new environment for capturing user interactions with the internet on PCs and Smartphones,

- speech recognition for new languages and new tools for mining audio streams,
- close to real-time analysis of contents and opinions in different languages and media,
- efficient technologies to support multilingual perception and multilingual business,
- a workflow for generating content analysis for different languages and media with minimal manually annotated data.

The following table indicates in detail the language analysis operators that will be applied or developed in SEMACORE.

| Language | Lang. Recognition | | Speech to Text | Machine Translation | KW Search & Disambig. | Content Classification | Opinion Mining |
|---|---|---|---|---|---|---|---|
| | Text | Audio | | | | | |
| Arabic | ✓ | ✓ | ✓ | + | + | + | |
| Catalan | ✓ | + | + | ✓ | + | + | |
| Dutch | ✓ | ✓ | ✓ | + | + | + | + |
| English | ✓ | ✓ | ✓ | ✓ | + | + | + |
| French | ✓ | ✓ | ✓ | ✓ | + | + | + |
| German | ✓ | ✓ | ✓ | ✓ | + | + | + |
| Italian | ✓ | ✓ | ✓ | + | + | + | + |
| Spanish | ✓ | ✓ | ✓ | ✓ | + | + | + |
| Turkish | ✓ | ✓ | + | + | + | + | |

**Figure 1 – Matrix of languages to be processed in SEMACORE with different analysis operators.**

**Operators marked by "✓" are already available for that language and will be applied and enhanced.**
**Operators marked by "+" will be developed and executed in SEMACORE.**

Language recognition is available for more than 60 different languages in text. For audio data language recognition is available for 28 languages[4]. The existing models will be updated for SEMACORE tasks and new models will be trained for the Catalan language. Speech-to-text models will be trained for Catalan and Turkish. Machine translation will be trained for Arabic, Dutch, Italian and Turkish. The baseline semantic analysis including keyword disambiguation, topic models, and classification will be performed for nine languages (Arabic, Catalan, Dutch, English, French, German, Italian, Spanish and Turkish). Opinion mining requires better linguistic resources and will be executed for six languages.

Additionally, to what is presented in the table, WP4 and WP7 will execute the following tasks:

- User data capture will be performed using web browsers and smartphones for collection, for all identifiable languages;
- Full web characterization will be performed in four countries: Netherlands, Republic of Ireland, Malta, and Luxembourg, for all identifiable languages;
- Web pages will be sampled with existing and proven method (Funredes) to serve as input for crosschecking and validation of other results for six languages (Catalan, English, French, German, Italian and Spanish).

[4] Bulgarian, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek, Hungarian, Italian, Latvian, Lithuanian, Polish, Portuguese, Romanian, Slovak, Spanish, Swedish, Russian, Arabic, Turkish, Urdu, Chinese, Hindi, Japanese, Korean

### B 1.1.4 Relevance to the topics addressed by the call

In the table below we summarize how the envisaged project contributes to the objectives listed in the FP7 -ICT Work Programme 2013, Objective ICT-2013.4.1, in particular for target outcome a):

> Innovative methods and tools for mining unstructured information embedded in text, speech, audio and video for the purposes of context-aware interpretation, correlation, aggregation and summarisation, turning information into usable understanding and actionable knowledge. Special emphasis is placed on social and collective intelligence from multilingual sources. Projects shall achieve broad coverage with efficient semantic interpretation. Of specific interest is the ability to capture sentiment and represent concepts and events, identify relations and similarities, interpreting time and space, within and across individual media, thus increasing our ability to detect and exploit otherwise hidden meaning across a range of applications.

We focus on expected outcomes and research themes as listed in the Call:

| Expected Outcome | SEMACORE |
|---|---|
| … mining unstructured information embedded in text, speech, audio and video | SEMACORE will crawl huge quantities of textual web pages, internet radio, podcasts, and video-sharing websites. The language of text and audio will be determined; audio will be transcribed to text, which in turn will be analysed semantically. |
| … social and collective intelligence from multilingual sources. | SEMACORE will perform most of its processing and semantic categorization analysis in nine major languages and opinions will be extracted in six languages. Sophisticated machine translation and topic models will ensure that the extracted features have identical meaning across languages. The analysis takes into account the relation between documents in social networks and will use this to improve semantic interpretation. By collecting user input from a panel of 10000 volunteers and analysing social network links we can take into account social relations and socio-economic features. |
| … ability to capture sentiment and represent concepts and events, identify relations and similarities, interpreting time and space, within and across individual media | SEMACORE will categorize documents with respect to genre and content. Keywords and aspects in a text coming from different media will be identified and disambiguated. Time stamps and location references and other names are provided by a Named Entity Recognition. Opinion mining extracts the relation between keywords, aspects and opinion words for text and audio coming from different media. |
| … achieve broad coverage with efficient semantic interpretation | For its use cases involving international organizations and companies SEMACORE will collect and analyse huge quantities of web data. If required semantic analysis will be performed online at near real-time speed. In addition the web of smaller countries will be crawled and sampled to determine actual language use and semantic content. |
| … context-aware interpretation, correlation, aggregation and summarisation, turning information into usable understanding and actionable knowledge | SEMACORE will provide an elaborate evaluation environment to combine the extracted features using a standardized representation and taking into account the relative uncertainties. Prominent evaluation dimensions are temporal and regional trends, links in social networks, and the socio-economic information on users available from the panel. The |

| | |
|---|---|
| | extracted information can e.g. directly be used to adjust political strategies or to restructure marketing campaigns. |
| … increasing our ability to detect and exploit otherwise hidden meaning across a range of applications. | The use cases address public organizations and enterprises and offer a unified access to keyword appearance and the opinion of users in different media, different languages extracted in near real-time. In addition our knowledge on the actual use of languages in the web is improved. |

### B 1.1.5 Timeliness of the proposed work

The web is the largest source of constantly renewed unstructured data and regularly used by 63% of the European population[5]. Scientific progress over the past twenty years has made great strides in extracting semantically meaningful knowledge from textual sources, and new services, such as Data as a Service are being proposed to exploit the richness of the web. Trend analysis, voice of the citizen, social concerns, e-Democracy, all these desires for better understanding involve exploiting the unstructured part of the web. A project such as SEMACORE is timely, since there is a need for being able to validate that the information extracted from the web is both complete and accurate. Moreover it is necessary to describe the contents produced and retrieved by user and to provide means to extract the opinions of citizens and customers. On the one hand this opens the opportunity to use information and communication technologies and strategies to ameliorate political and governance processes. On the other hand this offers the chance to improve products and services.

## B 1.2 Progress beyond the state-of-the-art

This section outlines the scientific and technological progress planned in SEMACORE. The technical details of the work packages are described in section B1.3.4.

### B 1.2.1 Beyond the state-of-the-art in dependable, real-time processing and storage for multimedia analysis

**State of the art:** Semantic analysis [Aggarwal 2012, Paass 2012] can be considered as a chain of annotation steps finally yielding the desired semantic interpretation. Each annotation step involves a model which is trained from comprehensive training data. Subsequently each trained model is applied in sequence to a huge stream of incoming documents adding annotations to the documents. To coordinate these tasks efficiently a runnable system is required to distribute the processing steps to many compute nodes. We will tackle this task with a Big Data architecture (lambda architecture [Marz 2012]) with separate batch and real-time layers. [Dhar 2012] argues that traditional database methods are not suited for knowledge discovery because they are not optimized for discovery of patterns in massive swaths of data. A specific requirement is to capture the uncertainty associated with models.
Specifically, our concern is in finding interesting and robust patterns that satisfy the data, where interesting is usually something unexpected and actionable, and robust means a pattern that is expected to occur in the future.
**Beyond state of the art:** The challenge is to apply current and possibly design new technologies of machine learning and model building on massive accumulated data sets based on Big Data concepts like map reduce technologies, or massive data streaming analysis with storage and retrieval of analytical result using NoSQL databases. In SEMACORE we will implement the lambda architecture containing a batch layer for processing stored data and a speed layer for processing streaming data. This will be extended for the analysis of unstructured data with specific workflows for training models on a

[5] Internet Users in Europe June 30, 2012. Internet World Stats, http://www.internetworldstats.com/stats4.htm

distributed MapReduce system. As an enhancement we will consider the Streams framework [Bockermann 2012] which defines a simple abstraction layer for data processing and provides a small set of online algorithms.
The predictions provided by different models are uncertain to some extent and errors propagate along the application chain. However, the uncertainties of models can be estimated in various ways, e.g. by bootstrapping [Davison 1997]. The SEMACORE system will support the administration of variance statistics and multiple models, e.g. for ensemble and bootstrap, to be able to estimate the uncertainties of predictions. It is most important, that for training and for application always the same workflow and the same parameters for pre-processing are used. Finally monitor workflows will be supported checking the deviation of statistics computed for new data. If the deviations get too large a revision of models has to take place.

### B 1.2.2 Beyond the state-of-the-art in user-centric measurement

The very concept of user-centric measurement as a whole is rather in. In its current interpretation of automatic data collection for statistical purposes, it was introduced by [Brennenraedts 2008] in an R&D project for Statistics Netherlands (CBS) and discussed in [Brennenraedts 2012]. The research design proposed for SEMACORE combines a browser-based (for desktop computers) and a proxy server-based (for mobile devices) approach, which has never been used before.

***Browser-based measurement***

**State of the Art:** Browser-based measurements are already widely applied, especially in online marketing research. However, language identification (if used at all) is usually limited to search terms and URLs and isn't used on the page content and text entered by the user.
**Beyond State of the Art**: In SEMACORE, we will link the URLs with an extensive database of websites that have already been characterized in terms of language. Using a browser extension, we can perform much of the analysis at the local level and mitigate several privacy issues by reporting only aggregate data. To further ensure (k-)anonymity, the front-end of the survey tool automatically tests whether the combination of the background variables does not lead to too small subgroups where unique individuals could in theory be identified [Sweeney 2002].

***Mobile-based measurement***

**State of the Art:** For SEMACORE, the mobile market is too important to be neglected due to its sheer size and important role in personal communication. However, on mobile devices, user-centric measures are usually restricted to a single app. App publishers are usually only interested in functional usage patterns, not language usage patterns. Platform owners such as Apple perform system-wide user-centric measurements, but only focus on functional aspects (such as the time that a particular application is open). Technical barriers (such as strict security measures and app sandboxing) prevent gathering system-wide statistics on mobile devices by parties other than the platform owner.

**Beyond state of the art:** In order to still be able to faithfully perform user-centric measurements on mobile devices, we have developed a novel method based on techniques borrowed from network-centric measurement. In network-centric measurement, the traffic sent and received from and to the mobile device is analyzed. By using this technology together with a new method to inject code for performing user-centric measurements inside the web pages opened by participants, we will be able to take a much closer look at language use than is possible using network-centric measurement alone. In addition, this method has the benefit that it is relatively cross-platform (almost all mobile web browsers are based on the WebKit browser engine, which means that we can support many different mobile operating systems while still targeting a single engine).

Another shift in the current technology frontier is the combination of the online client with an online survey module. Focused survey questions can be triggered by specific events (that is, online usage patterns). Although such event-driven online surveys applications are already commercially available,

the proposed method is more efficient and targeted, as the events are based on actual language use (and not on basic indicators, such as a particular URL being visited).

To safeguard privacy, an innovative heuristic will be used that is built on the off line sampling heuristic from [Van Alstyne 2003]. The heuristic enables the recognition of specific words (which function as language markers) without revealing the actual semantic content of the personal documents. The proxy involved in the measurements for mobile devices will be run at the lowest possible level (which may even be on the mobile device itself).

### B 1.2.3 Beyond the state-of-the-art in spoken language analysis

***Language identification from speech***

**State of the art:** There are two main approaches used by language recognition systems: acoustic and phonotactic. Acoustic approaches rely on the acoustic parameters derived directly from the speech signal [Lamel 1995], and modelled with discriminatively trained Gaussian Mixture Models [Burget 2006] or with Support Vector Machine-GMM Super Vector [Singer 2012]. The performance of acoustic approaches has improved significantly using a variety of channel compensation techniques [Kenny 2007, Castaldo 2007]. Phonotactic approaches rely on the assumption that the sequences of phonemes, that is how the sounds follow on another in words and sentences, are language specific [Harper 2008]. This means that even if two languages share the same set of phonemes, their phonotactic characteristics are different. Phonotactic approaches typically use one (Phone Recognizer followed by Language Modelling (PRLM) approach) or multiple (Parallel Phone Recognizers followed by Language Modelling (PPRLM) approach) well trained phone recognizer(s) [Zissman 1996] to derive phone n-gram statistics. These phonotactic characteristics are then used to estimate models for language recognition. Today's best performing language recognition systems combine both phonotactic and acoustic sub-systems [BenZeghiba 2009, BenZeghiba 2012, Brummer 2012].

**Beyond state of the art:** The performance of the widely used phonotactic approaches is highly dependent on the quality of the underlying phone decoders. Therefore one research direction will be to improve the underlying language-specific (L) and language-independent (LI) phone recognizers in order to generate more consistent phone n-gram statistics. In SEMACORE, work in WP5 will investigate several techniques that have been widely adopted in speech recognition but have not yet been applied to language identification. These include discriminant training for the component acoustic phone models; methods to improve the quality of target language phonotactic models such as optimization of decoding parameters and intelligent selection of phone contexts; and improved decoding making use of multiple hypotheses and automatic learning techniques.
Another challenge that will be addressed is dealing with multilingual audio documents. Today's state-of-the-art language identification (LID) systems typically assume that an audio document is only in a single language. This assumption is not always valid, in particular when there is simultaneous translation of audio segments containing speech in a language other than the main language of the document [Liu 2012a]. To address this LID will be compared on fixed sized second chunks and on clustered segments found by an automatic partitioner.
For the most part language identification systems have been developed for the classification of telephone speech, and evaluated in regular benchmarks run by NIST[6]. In WP5 of SEMACORE, state-of-the-art techniques will be adapted and improved so that they can be successfully applied to the heterogeneous audio data found on the web.

***Automatic speech recognition***

**State of the art:** Automatic speech recognition, also called speech-to-text transcription, is concerned with converting the speech waveform, an acoustic signal, into a sequence of words. Today's best performing approaches are based on a statistical modelling of the speech signal. The CNRS and VOCAPIA

[6] US National Institute of Science and Technology (NIST) www.itl.nist.gov/iad/mig//tests/lre

Research have collaboratively developed speech-to-text transcription systems for broadcast and web data in 15 languages. Automatic speech recognition has been the focus of many technology benchmarks, both in the US (NIST sponsored evaluations http://www.itl.nist.gov/iad/mig/tests) and Europe (SQALE [Steeneken 1995], ESTER, NBEST[7]). Today's transcription systems are typically trained on huge heterogeneous audio and text corpora.

**Beyond state of the art:** In the SEMACORE project the CNRS and Vocapia will apply novel unsupervised methods to train both acoustic models and linguistic models for the Catalan and Turkish languages, in order to support further downstream processing [Lamel 2002, Lamel 2010]. Turkish being an agglutinative language, ensuring high lexical coverage (that is the words known by the system) poses challenges that will be addressed by automatic discovery of lexical units**.**

### B 1.2.4 Beyond the state-of-the-art in content analysis

***Statistical Machine Translation***

**State of the art:** Statistical Machine Translation (SMT) is nowadays the dominant paradigm in Machine Translation (MT). From the pioneering works on statistical machine translation by [Brown 1990, Brown 1993], the field has experienced several enhancements. It was soon noticed that translation should not be reduced to a word-to-word process. The contextual information helps the translation process, making it convenient to consider translations between *n* consecutive source words into *m* consecutive words in the target language. This fact motivated the usage of phrases as translation units and consequently the birth of phrase-based SMT [Och 2004, Koehn 2003]. Further enhancements involved the integration of more abstract levels of linguistic information in the translation process, instead of working with word strings alone. This is the case, for instance of factored translation models [Koehn 2007], which allow to deal with sequential information naturally (lemmas, morphological features, etc.), or hierarchical or syntax-based SMT models, which allow syntactic structure to take part in the source-target alignments and in the translation process [Yamada 2001, Chiang 2005, Marcu 2006, Liu 2008, Zhang 2008, Carreras 2009]. Recently, SMT approaches incorporating rich linguistic information started to perform significantly better than pure phrase-based SMT, and represent a very active focus of research in the MT community. SMT will be used in SEMACORE to port training documents in languages for which no such resources are available.

**Beyond state of the art:** The main challenges related to MT that will be addressed in SEMACORE are those related to genre and domain adaptation for Machine Translation systems. We will research on parameters tuning techniques to adapt a model trained in a general domain corpus (e.g. EuroParl) in order to allow it to operate in different text types (e.g. web pages) or semantic domains. Depending on the available corpus for the target domain (parallel, comparable, monolingual,...) different aspects of the statistical models can be adjusted.

***Keyword Disambiguation***

**State of the art:** Current linguistic approaches to word sense disambiguation mainly rely on word similarity distances gained by evaluating parallel corpora [Lefever 2010]. An alternative uses multilingual topic models which assume a bag-of-words approach. They are trained using documents with roughly the same contents across languages (e.g. Wikipedia articles in different languages) and assume that these documents share the same tuple-specific distribution over topics [Mimno 2009]. Recent approaches in addition exploit a hierarchical thesaurus, e.g. multilingual WordNets to derive topic models [Boyd-Graber 2010], which are, however, restricted by the limited coverage of WordNet. Fraunhofer IAIS has extensive expertise in large scale topic models [Wahabzada 2010, Wahabzada 2011], developed an approach to detect supersenses from WordNet with Conditional Random Fields [Paass 2009] and has used additional information for disambiguation extracted from entities occurring

[7] ESTER http://www.afcp-parole.org, NBEST http://hmi.ewi.utwente.nl/project/STEVIN

in the neighbourhood of expressions [Pilz 2012]. The last approach uses the efficient implementation of a search index.

**Beyond state of the art:** In SEMACORE we will train multilingual topic models based on parallel word-aligned documents [Och 2003] generated by human and machine translators, which provide much more evidence on the correspondence of words in different languages than roughly corresponding Wikipedia articles. Using human-generated parallel texts in addition avoids the errors introduced by machine translation. As topic matching between different languages usually is not possible at the word level we will match topics simultaneously at the document, sentence, and entity level. In addition we will include available online multilingual resources (dictionaries, linked open data, etc.). Similar to [Pilz 2012] we will extend the multilingual topic models by collective disambiguation for multiple keywords/entities and exploit them for disambiguation.

### *Keyword Search*

**State of the art:** The search of keywords in multilingual resources has been investigated in Cross-Language Information Retrieval (CLIR). [Sorg 2012] reviews the state of the art and utilizes a multilingual topic models based on Wikipedia.

**Beyond state of the art:** We will use our extended multilingual topic models based on parallel text to disambiguate candidate hits. Concerning content search in audio, speech processing technology will be exploited to provide the possibility to search and select audio and video documents directly from their contents. This is achieved by using automatic transcription and topic/content detection methods to annotate the unstructured data sampled from the web. Textual as well as extra information obtained from speech may be included the topic model.

### *Text Categorization*

**State of the art:** Category recognition aims at classifying web documents using a hierarchy of predefined content categories [Bi2011]. The SEMACORE context poses a completely new challenge, requiring content classification and opinion mining for many languages, where the content categories need to be identical across the languages. Advanced algorithms typically rely on annotated training data which are not available except for a few languages. There is an alternative approach to classification using topic models [Rubin 2012]. [Ni 2011] extend this to multilingual topic models based on Wikipedia. For named entity recognition we will evaluate novel multilingual approaches using Wikipedia labels in a semi-Markov CRF [Kim 2012]. Fraunhofer IAIS analysed Twitter messages and extracted spatio-temporal patterns to describe the regional and topical messaging habits of people [Andrienko 2012].

**Beyond state of the art:** In SEMACORE we will have access to multilingual topic models based on parallel word-aligned documents [Och 2003]. In addition we will translate extensive training data and additional documents annotated with existing classifiers from resource-rich languages to other languages using machine translation. Finally the training documents will be annotated with cross-lingual topics. Based on this rich feature set we will train new classifiers, e.g. using SVMs. As an alternative we will extend our multilingual topic models directly for classification [Rubin 2012] and compare them to previous classifiers. In the same way multilingual training data for named entity recognition with CRFs will be generated. Specificities for spoken language will be addressed, considering different configurations for speech recognition and different outputs (word lattices, confusion networks or N-best lists, confidence scores, word classes ...).

### *Opinion Mining*

**State of the art:** Opinion mining [Liu 2012] (also called sentiment analysis) focuses on the automatic identification of subjective expressions that describe people's sentiments or feelings toward entities, events and their properties, e.g. as positive, neutral, negative. Actual methods among others are based on Conditional Random Fields (CRF) [Jakob 2010, Paass 2012a] and parse trees [Wu 2011]. Fraunhofer developed the lifted inference approach to estimate extremely large sized relational models [Neumann

2011, Ahmadi 2012, Kersting 2012]. [Banea 2011] provides an overview on cross-lingual opinion mining and advocate cross-lingual projections of training data. They show that the inclusion of multilingual information can improve monolingual sentiment analysis [Banea 2010]. [Socher 2011] introduces a monolingual deep learning model which is trained to predict the words of a sentence through a bottleneck (autoencoder). As the resulting latent representation of words explicitly takes into account the structure of a sentence [Collobert 2012] it can extract the relation between an aspect and the sentiment in a much better way than a topic model approaches based on a bag of words representation.

**Beyond state of the art:** To arrive at a model for the opinion with respect to specific keywords we will extend deep learning models to a multilingual setting by exploiting aligned parallel texts in multiple languages, possible enhanced with topic multilingual topic distributions. First a number of deep learning auto-encoder models will be trained in different languages where the training corpora include parallel aligned corpora (e.g. the Europarl-corpus) and aligned translated training data[8]. After the training of the auto-encoders in different languages the opinion classifier is trained explicitly on the labelled data. On the unlabelled parallel data the opinions are trained to be identical for both languages, thus extracting additional evidence. In contrast to standard topic models this deep learning model will include sequential structural information on the different target languages and in addition evidence from unlabelled samples. It will be compared with classical multilingual approaches. We will also evaluate if opinion mining can benefit from then information contained in parse trees based on our experience with parse-tree based relation extraction [Reichartz 2009, Reichartz 2010]. Finally we will investigate if opinion mining based on a translation to a resource-rich language (e.g. English) can improve performance.

For audio data, opinion analysis will be based on statistical models trained on pre-selected data associated with the transcription. Based on the output of the speech-to-text system, linguistic analysis will allow the detection of opinion-carrying words. A paralinguistic analysis based on measures in the speech signal, such as the speaking rate, energy, pitch, as well as speech fluency (hesitations, pauses) frequency, will provide additional information to give an indication of the observable opinion or sentiment of the speech. The information carried in the audio and in the automatic transcription will be fused and may be included in the deep learning models.

### B 1.2.5 Beyond the state-of-the-art in using the analysis of languages, content and opinions

***Measuring Linguistic Diversity (Use Case 1)***

**State of the art** : The theme of characterizing the web and, in particular, measuring the linguistic diversity on the Internet, which remained, prior to 2006, the reserve of a small group of specialists has been practically stuck afterwards, in spite of the growing interest of several international parties (such as ITU [ITU 2009] or UNESCO [Paolillo 2005]). The unique updated source as of today[9] is limited to Internet users per language for only the top 10 and with a limited methodological trust. No consistent series of data exist anymore on repartition on languages in the web, only partial and dispersed data for some applications can be collected (such as, for example, for Wikipedia[10], Twitter[11] or Facebook[12]).

UNESCO has published a report [Pimienta 2009] compiling the different approaches and evaluating future prospects based on experience accumulated during twelve years of research in the subject area. At a time when interest in the theme is becoming universal, the work of the pioneers is experiencing

[8] e.g. MPQA corpus (http://cogcomp.cs.illinois.edu/Data/MPQA_data/) or NTCIR-EN corpus (http://research.nii.ac.jp/ntcir/data/data-en.html) as well as use-case specific training data
[9] http://internetworldstats.com/stats7.htm
[10] http://stats.wikimedia.org/EN/Sitemap.htm
[11] http://bigthink.com/strange-maps/539-vive-le-tweet-a-map-of-twitters-languages
[12] http://www.insidefacebook.com/2010/05/24/facebooks-top-ten-languages-and-who-is-using-them/

difficulties as a result of developments in the Net and search engines[13]. Accordingly, there have been no productions to consult since 2007, when two MAAYA members, LOP[14] and FUNREDES[15], the two most visible actors in producing indicators, published their most recent works.

The methodology used by LOP [Suzuki 2002] consists of systematically crawling all the pages of domains of countries to be studied and identifying their script in order to count the pages in a given language[16]. Where a script is shared by a number of languages (as is the case of the Western languages), a language recognition algorithm is applied using variable n-gram order [Choong 2009]. The method reaches its limits though when it comes to focusing countries with a vast amount of pages, such as China and Korea and, for the same reasons, is not addressing large generic domain names (such as .com, .net or .org). Furthermore, the growing trend in managing country code top level domains (ccTLD) is to allow its utilization by entities remote to the country, for business purposes, and this may provoke growing biases in the results.

The methodology used by FUNREDES, limited to a number of Western languages[17], is based on the selection of a vocabulary in these different languages with appropriate characteristics in terms of equivalence, range and cultural neutrality. The counting by search engines of the pages corresponding to this vocabulary enables their respective percentages to be compared with statistical tools. This has allowed the largest and more stable history of measurements since 1998, including interesting results by countries or for other Internet spaces. This approach is, however, no longer reliable since the counters offered by search engines cannot be trusted anymore. In addition, it is no longer feasible to extrapolate the results to the entire universe since the space indexed by the engines represents now a far smaller proportion of the total space[18] (and especially as linguistic bias has appeared, as a consequence, in the sample indexed).

The very nature of the web has evolved, transforming it into a dynamic and infinite space, and there have been no more attempts to characterize such a huge and moving target. However, more than ever such data is required for informed decision making by many stakeholders, from policy makers tackling the digital divide[19] to industrial players of the digital economy willing to establish reliable business cases for their ventures and understand the evolution of the markets.

**Beyond state of the art**: In this context, SEMACORE will make a breakthrough in a field where the state of the art has been regressing, by putting again web characterization and language measurement in the research agenda and by expanding the potential for building indicators for spaces and formats other than the static and text-only web. Additionally to language detection SEMACORE will determine content categories, genre, and/or sentiment to reveal which content is communicated in what language. As for the current limitations on crawling, new approaches such as stream sampling [Nesreen 2012] will be applied. In contrast to previous approaches we will also extract the language of audio sources in

[13] On the one hand, the size of the web makes the work of systematically browsing pages ("crawling") ever more problematic (we could say that is an infinite space), and on the other, partially as a result of the first point, search engines only index an increasingly insignificant percentage of the visible Web and the indications offered on the total number of occurrences of the words searched are no longer credible at all (even though some methods are based on them).

[14] Language Observatory Project - http://www.language-observatory.org/

[15] http://funredes.org/LC

[16] LOP after having focused local languages in the Asian and African Web is conducting similar studies in Latin America and the Caribbean in collaboration with FUNREDES.

[17] Catalan, English, French, German, Italian, Portuguese, Spanish and Romanian.

[18] This percentage has fallen from 80% until 2004 to less than 30% in 2007 to less than 10% as of today.

[19] While historically the efforts to overcome the digital divide has been essentially focused on providing more physical access, a new trend is prone to arise, to focus the content divide which is linked to digital literacy and linguistic diversity and which figures provided by NUT and FUNREDES have shown it is an order of magnitude deeper than the access divide,

combination with content analysis. Finally we will compare the results of previous approaches with that of new methodologies to assess the gain in explanatory power.

***Measuring web content and opinions for an international public agency (Use case 2).***

**State of the art**: Sentiment analysis for public agencies attempts to identify the sentiment or attitude in a text span on a public or political issue. Analyzing publicly available data to infer population attitudes is faster and less expensive than traditional polls and can actually provide a more "real-time" view of the current political climate [O'Connor 2010]. In a review [Malouf 2008] conclude that it is quite difficult to extract the political attitude of citizens towards vague concepts, e.g. "more democracy". They also suggest to use the analysis of the user's social network to improve analysis. A state of the art description is given by [Osimo 2012]. [Mukherjee 2012] captures vague concepts by a topic model. Nevertheless political opinion mining is quite challenging, because few political issues can be described with one or two words. Political sentiments are also harder to determine due to complex mixture of factual reporting and subjective opinions, and heavy use of sarcastic sentences [Liu 2012].

**Beyond state of the art:** In SEMACORE we will concentrate on opinions expressed with respect to some name or keyword, e.g. "smoking", as this leaves less room for misinterpretation. We will apply advanced opinion mining techniques developed in WP6 to extract opinions concerning an actual issues expressed by a keyword. Moreover we will identify aspects of keywords, e.g. "ban" for "smoking" and determine the opinions on these aspects. Aspects will be grouped semantically by topic modeling and similar tools. As in addition the content / genre category of the web document will be analyzed and as multiple opinions may be expressed in one document (e.g. "I like smoking", "I think a smoking ban is good") many interesting facets may be revealed which are highly relevant for the customers and decision makers. Additional evidence will be gained by including named entities in the analysis, e.g. disease names or cigarette brands. By analyzing the dynamics of disease mentions and their symptoms (and the opinions on that) it is even possible to detect an outbreak of an epidemic [Paul 2011].

***Measuring web content and opinions for an international company. (Use case 3).***

**State of the art**: There is a growing literature on opinion mining with respect to products [Liu 2012]. Especially popular is the analysis of review web sites, e.g. Amazon, where reviews are grouped with respect to each product. Usually there are summaries for each review which contain an overall judgment with respect to a few product features, e.g. durability, performance, or handling. As [Liu 2012] states, lexicon-based approaches with sentiment words can handle about 60% of the cases. Using pattern-based algorithms is difficult, because there are simply not enough training data for patterns. In addition opinions are expressed in domain-specific ways. Therefore it is important to concentrate on a specific domain.

**Beyond state of the art:** In SEMACORE we will concentrate on a domain for a company to adapt our advanced multilingual approach for mining opinions with respect to keywords. We will employ topic and sequence information to identify the target of the opinion and take into account negation and modality to determine the opinion orientation [Benamara 2012]. We will apply advanced opinion mining approaches developed in WP6 to extract opinions concerning a company, product or issue expressed by a keyword and its aspects. We will exploit the sequences/links of posts (e.g. on Twitter) to enrich the extraction of single documents. Additional information will be provided by extracted named entities, e.g. product and company names. A major issue is the data noise, especially all kinds of spelling, grammatical, and punctuation errors, which we will tackle with cleaning techniques [Dey 2008]. A flexible retrieval and evaluation environment will be provided to allow ad-hoc queries, e.g. on the relation between opinions on different aspects. For aggregation and presentation of results, e.g. in dynamic semantic maps, we will analyze the available tools [Osimo 2012] and select the most appropriate dynamic and interactive procedures.

# B 1.3 S/T methodology and associated work plan

## B 1.3.1 Overall strategy

The work is organized in a number of technical work packages (WP4-6) as well as a work package for use cases and evaluation (WP7). Additional work packages are dedicated to project management (WP1), to legal issues and linguistic diversity (WP2), and to exploitation (WP8).

| | Title | | Lead |
|---|---|---|---|
| WP 1 | Project Management | 1 | ERCIM |
| WP 2 | Societal Issues | 2 | MAAYA |
| WP 3 | System Development & Integration | 7 | FRAUNHOFER |
| WP 4 | User-centric Measurement | 4 | DIALOGIC |
| WP 5 | Analysis of spoken language in multimedia data | 5 | CNRS |
| WP 6 | Content analysis on multimedia data | 3 | UPC |
| WP 7 | Use Cases | 10 | NIELSEN |
| WP 8 | Dissemination & Exploitation | 2 | MAAYA |

**Figure 2 – Overview of work packages**

SEMACORE will implement its work plan over three years, with major deliveries at M12, M24, and M36.

## B 1.3.2 Timing of the different WPs (Gantt)

| | | | Year 1 | | | | | | | | | | | Year 2 | | | | | | | | | | | Year 3 | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| | | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 |
| **WP 1** | **Project Management** | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | **T1.1** | Co-ordination of activities | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | **T1.2** | Resource management | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | **T1.3** | Project reporting | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | **T1.4** | Deliverable quality assurance | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | **T1.5** | Collaborative tools | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| **WP 2** | **Societal Issues** | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | **T2.1** | Intellectual property matters | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | **T2.2** | Legal, ethical and regulatory matters | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | **T2.3** | Privacy protection and security technologies | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | **T2.4** | Linguistic diversity & information society indicators | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| **WP 3** | **System Development & Integration** | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | **T3.1** | Scientific coordination | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | **T3.2** | Terminology Definition and System Requirements | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | **T3.3** | System architecture | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | **T3.4** | Project infrastructure, processes and tools | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | **T3.5** | Prototype development cycle | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| **WP 4** | **User-centric Measurement** | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | **T4.1** | Input specification: indicators, languages. panel, survey | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | **T4.2** | Supported configurations specification | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | **T4.3** | Panel management | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | **T4.4** | Security and privacy audits | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | **T4.5** | Development of the desk-top based user-centric measurement core client | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | **T4.6** | Development of the desk-top based user-centric measurement (survey module) | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | **T4.7** | Development of user-centric measurement for mobile devices | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| **WP 5** | **Analysis of spoken language in multimedia data** | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | **T5.1** | Specification of the representative corpus obtained via large scale sampling | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | **T5.2** | Annotation of the representative corpus with language tags | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | **T5.3** | Language identification from audio | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | **T5.4** | Model Development for Automatic Content Category Identification | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| **WP 6** | **Content analysis on multimedia data** | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | **T6.1** | Intelligent data gathering and language identification | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | **T6.2** | Prepare multilingual training data | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | **T6.3** | Disambiguation and classification | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | **T6.4** | Opinion mining | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | **T6.5** | Evaluation and aggregation of results | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| **WP 7** | **Use Cases** | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | **T7.1** | Linguistic diversity use case | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | **T7.2** | Content and opinion mining for an international public agency | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | **T7.3** | Content and Opinion mining for an international company | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | **T7.4** | Evaluation of language indicators | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | **T7.5** | Use cases evaluation and future impact forecast | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| **WP 8** | **Dissemination & Exploitation** | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | **T8.1** | Web site and production of promotional material | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | **T8.2** | Institutional exploitation and sustainability roadmap | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | **T8.3** | Academic and scientific dissemination | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | **T8.4** | Commercial dissemination and Exploitation | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |

Figure 3 - Gantt chart

## B 1.3.3 Components of the different WPs (Pert)

### *Work package and tasks dependencies*

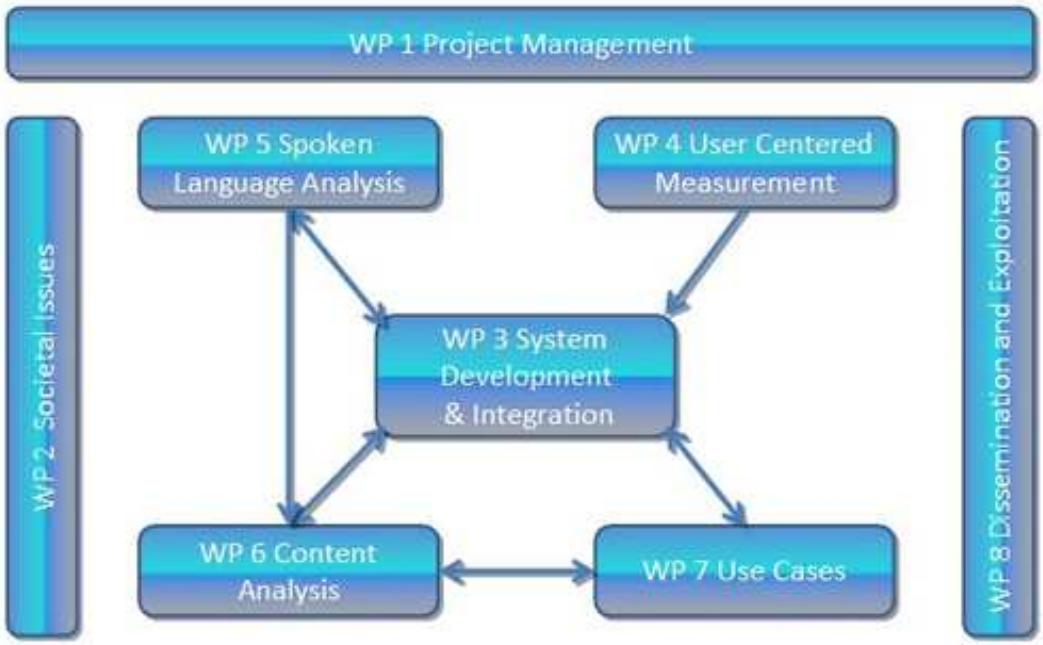


**Figure 4 - Work package dependencies**

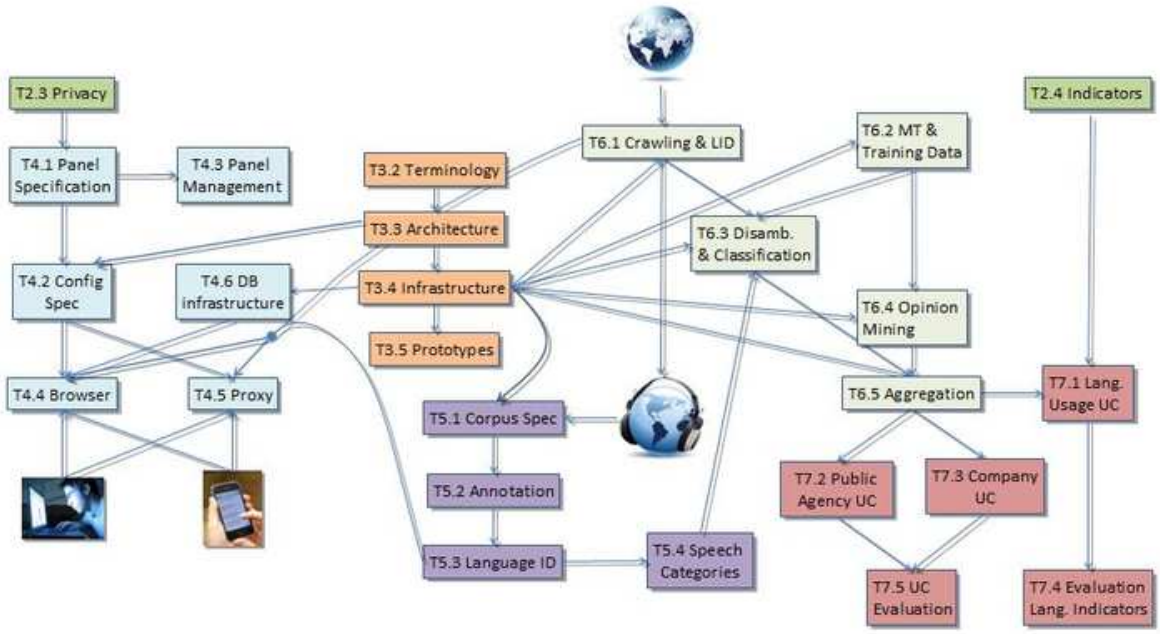


**Figure 5 - Task dependencies**

## B 1.3.4 Detailed description of work packages

SEMACORE work can be divided into three parts: technical work on collecting multilingual data and extracting useful semantic information (WP4, WP5, WP6); applicative work using this data (WP7); and dissemination and management (WP1, WP2, WP8). Here is a description of the non-management work packages.

### *Technical work package overview*

### *WP3 - System Development & Integration*

Summary:
We develop an infrastructure and platform that coordinates technical modules developed in WP4 through WP7. It can process huge amounts of unstructured data in near real-time with guaranteed reliability, fault tolerance and distributed computing. It allows specifying flexible workflows which integrate data collection, data storage, data analysis and evaluation. The platform is a framework for the successive integration of separate modules into increasingly comprehensive prototypes.

This work package provides a runnable system, that integrates the diverse technologies into a framework according to a Big Data architecture (lambda architecture) with a batch and real-time layer.

The following figure shows an instantiation of this architecture using open source components. These open source components are a typical selection of well known products. There are alternatives and within the first year of prototype development a selection that most appropriately meets our challenges will be evaluated and selected.

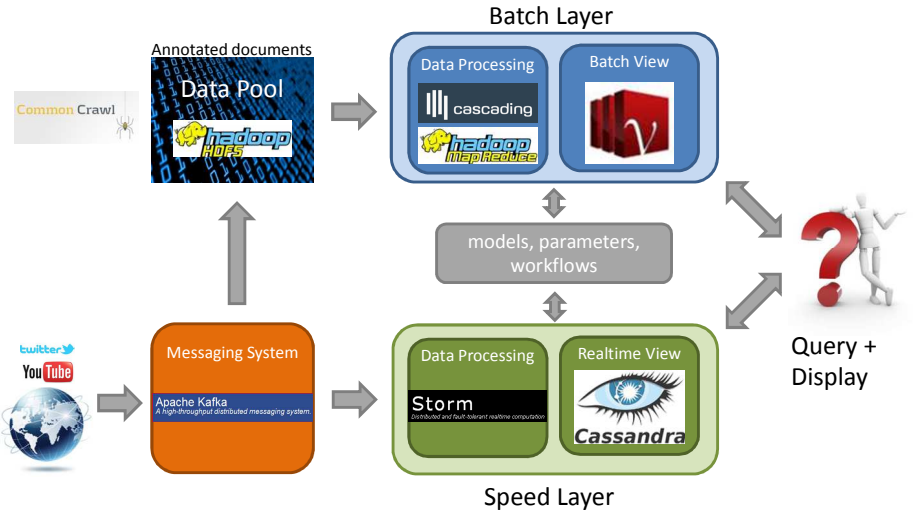


**Figure 6 - SEMACORE Open Source Architecture**

The messaging system collects data from the internet and could be implemented using Apache Kafka[20]. The data pool is a repository for downloaded and annotated data and will be realized in a distributed manner, e.g. using Hadoop hdfs[21]. The batch layer is designed to process extensive data on a network of processors. Learning and training components will be integrated into the batch layer utilizing distributed processing with Hadoop and Cascading[22] to create and execute complex data processing workflows on a Hadoop cluster. These workflows create new trained models and parameters that subsequently can be

[20] http://kafka.apache.org/
[21] http://hadoop.apache.org/docs/r0.17.1/hdfs_design.html
[22] http://www.cascading.org/

applied in the speed layer to the incoming stream of data. The speed layer is designed for processing massive streams of incoming data in near real-time and may use Twitter Storm with Apache Cassandra for handling the data[23] .

The separation of a batch and a speed layer within the lambda-architecture allows to separate time-consuming offline analysis tasks, for example the training of machine learning methods, from real-time interpretation of incoming data. A separate workflow system explicitly handles the coordination between the layers. The SEMACORE system will be designed to support near real-time analysis. This covers the process of crawling, pre-processing, analysis and aggregation of results. For some use cases this will be important, e.g. live covering of the opinions expressed in an event (e.g. an election). Problems may arise for speech recognition which can process speech in real-time but currently is file oriented. By using distributed processing we will be able to get results in a time span below a minute, which is sufficient for our use cases.

The system is developed in a series of prototypes, three of them are accentuated as stable system releases to be provided for evaluation in the context of WP7 use-cases and for presentation at the annual reviews. The infrastructure for collaboration will be set up as soon as possible; the first prototype according to the lambda architecture will be available within the first six months, thus providing the platform for experiments and evaluation.

Development operates with agile sprint and Kanban methods. A collaborative workspace (Atlassian Confluence wiki) for the secure exchange of project information, project results, deliverables, meeting minutes, etc. facilitates easy and timely inter-project communication, in addition to mailing lists. An issue tracking system (Atlassian JIRA) is employed for the management of service development and deployment work items. Technical requirements and development steps are selected and documented in close cooperation with WP4, WP5, WP6, and WP7 partners. The system and requirement reports of the stable prototypes are collected and summarized for the documentation and for the issue descriptions in the collaboration platform.

The system requirements are mainly determined by the use-cases in WP7 and can be developed further and checked against the evolving system, especially at the steps of the stable release versions.

### *WP4: User-centric measurement*

Summary:
For each client, information on which language is being used in which type of web-based application will be collected at a particular moment in time (day time, evening, working days, weekends) and a particular frequency. Both language use on desktop PCs and mobile devices is being monitored, the first via a browser extension (including a survey module that is used to validate the data that is automatically collected via the core client), the second via a local proxy server. The data is unique in its kind, and enables in-depth analysis of language-in-actual-use across specific background variables (gender, age, nationality etcetera), and across specific types of use settings (type of online activity, time).

This WP measures language behaviour at the most detailed level and at the point of origin: the individual user. Both the language use on desk top PC's and mobile devices will be measured albeit in distinctively different manners. (1) The desk top based measurements use a sophisticated piece of software that is installed as an extension to the web browser. (2) The measurement at mobile devices uses a local proxy server over which all outgoing IP-traffic from the devices is being routed.

[23] Twitter Storm: https://github.com/nathanmarz/storm; Apache Cassandra: http://cassandra.apache.org/

***Desktop based user-centric measurement***

The browser extension (1) is a script that automatically identifies the languages that are being used by the end user on that particular device, in various settings.[24] It basically covers any text that is being sent via the web browser.

Giving the highly sensitive nature of the data that is collected by the client, maximum care will be given to the privacy of the respondent and the security of the data. We have dedicated a specific task (4.4) to a strictly independent party (Kyos) for security and privacy audits of both (1) and (2).

Users have full control over the privacy settings. For the online tracking behaviour, for instance, they can block specific websites for being tracked, or temporarily pause tracking.

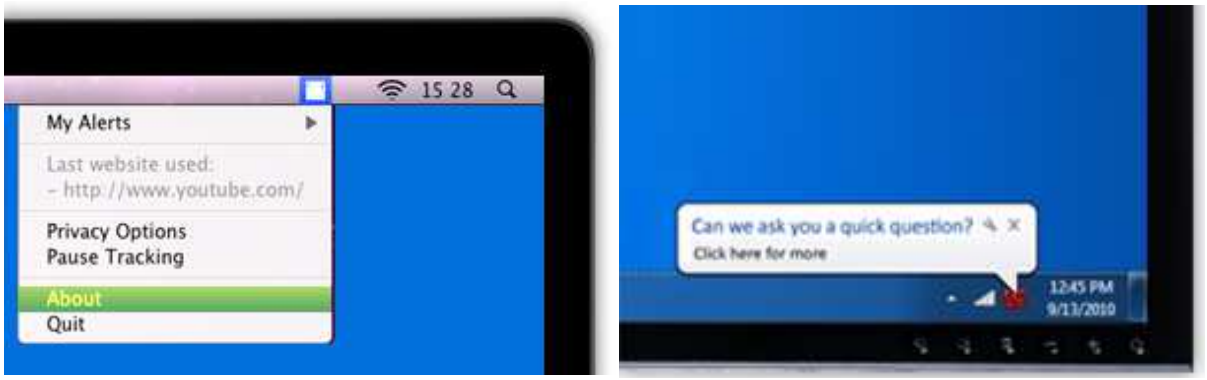


**Figure 7 - Dummy screenshot of the online module (left) and of the event-driven survey module (right).**

All data will be anonymised locally, at the client side, before it is being sent over an encrypted connection. Only data at a meta-level of the original texts (e.g., type of language, generic characteristics of the text etcetera) will be transferred to and stored at the central server. Consequently, the language identification module will run at the client side, not at the server side, and should therefore be extremely light and fast to minimize use of resources from the devices on which the client is being installed.

The core client also has a built-in survey module.[25] Survey questions appear in pop-ups in the client and can be triggered automatically, by specific events, or manually.

In the first case, the questions and triggers are programmed in advance by the researchers but triggered by specific events in the automated agent. In the latter case, tailor-made questions are sent to specific subsets of respondents. The identity of the respondent is at no time known or revealed. The ID in our micro data is matched to a specific respondent via hash tables. Thus we can send survey questions to a specific respondent without ever knowing the identity of the respondent.[26] We will develop a web-based interface that enables user-friendly access to the (large amounts of near-real time) data. The interface is used to link the input and output from the core and survey module on the level of individual (anonymous) respondents or subsets of respondents. This means that we can immediately follow-up specific patterns of language behaviour by specific groups of users with specific questions (e.g., to clarify the behaviour and/or asking more detailed information or types of variables that cannot be covered by non-intrusive automated data collection, such as perceptions). One particular event that always triggers the survey module is the initial installation. After installation of the client some questions will be asked

[24] We propose to develop tailor-made clients for the most popular web browsers, Microsoft IE (7.0 to 10.0), Google Chrome, Apple Safari (4.0-6.0), and Mozilla Firefox (4.0 and up)

[25] Use of the survey module is strictly limited to the purposes of this research project. The most important goal of the survey module is to directly collect data from respondents that can be used to validate the data that is automatically being collected about the respondent's behaviour.

[26] To further ensure (k-)anonymity, the front-end of the survey tool automatically tests whether the combination of the background variables does not lead to too small subgroups where unique individuals could in theory be identified (Sweeney, 2002).

to create a personal profile for each household member (that is, each potential user of the desk top computer). Secondly, several questions will be asked to fill the personal profiles.[27]

Although the client only collects a modest amount of data for each unique language action of the end-user (a string of characters), the total amount of data generated is still very big due the very high frequency of the measurements. In previous pilot studies, with a relatively small panel size of 10,000 active users, the client already generated 250 megabytes of data on a daily base, or 7.5 gigabytes per month [Brennenraedts 2012]. Note that this measurement campaign only covered a small part of the online behaviour, namely the URL's in the browser cache. The proposed measurements in this project have a much wider scope (e.g., include all online behaviour and also offline behaviour) and contain much more details. Hence the input files and processed files will be several magnitudes (50-150 times) bigger than in the previous pilot. We estimate the total monthly amount of data that will be generated between 0.4 and 1.1 terabytes.

***Mobile device based user-centric measurement***

For the mobile device measurement (2) we have chosen to use a proxy server because this basically renders us independent of the type of device and type of operating system. Otherwise we would have had to develop dedicated clients for every device/OS-combination – and there are hundreds of them in the market. Instead we use a proxy server to analyse all outgoing IP-traffic. The traffic will be only temporary be stored in a buffer, be analysed for language use, and then be deleted. This technical solution is feasible for mobile devices because the average traffic flow per user is still relatively small compared to fixed broadband.[28]

One major challenge in this task is the initial set-up of the proxy server in the default settings of the mobile device. This is still very much platform and OS-dependent. Hence we will have to come up with a tailor-made solution for every device that we want to include in the measurement.[29] The latest versions of iOS does not permit the use of global proxies (earlier versions do not seem to have this limitation) whereas the latest versions of Android even support running a proxy on the device itself (but not the earlier versions). Hence it is critical to test the feasibility of the various configurations. Due to the challenges in setting up the local proxy servers we have chosen to run the mobile measurements in a limited number of selected Member states, namely the Republic of Ireland, Luxembourg, Malta, and The Netherlands.

As far as we know this is the first time that actual online language behaviour of users will be measured on a large (EU-wide) scale, across a variety of countries, languages, user background variables (gender, age, nationality etcetera) and user contexts (type of online activity, time), on a near real-time basis. The combination from the survey modules – questions that are automatically triggered by events in the automated language identification agent – is also highly innovative. The survey instrument could also be used in a more generic manner, to measure many other dimensions as well (e.g, opinions about specific topics, see Task 6.4).

### *WP5 - Analysis of spoken language in multimedia data*

Summary:

[27] Some of these questions related to the aforementioned background variables (which are usually not known ex ante). Some additional questions will cover the complementary usage of electronic devices (that are not covered by the measurement). To improve the external validity of the measurement results, it is important to know which part of the relevant activities is not covered by the measurement. We can also derive some useful background information from the language settings in the operation system.

[28] Maximum 100MB per month.

[29] We propose the following configurations for the feasibility tests: iOS (4 to 6), Android (3.*,4.*), Blackberry (5, 6, 7.*), and Windows Phone (7.*, 8.*)

Work package WP5 is concerned with developing methods to identify the language in the diverse types of audio data found on websites hosting audio and video documents. As a lot of the information on the web is not in a textual format, these cannot be detected and categorized via text-based methods. A subset of representative data will be selected in coordination with the other work packages, mainly WP6 and WP7, and annotated with language labels in WP5 (task 2). An additional challenge that will be addressed is the recognition of highly representative language variants. In coordination with the text-based methods developed in WP6, Task 5.4 will explore techniques to identify the topics and other information in audio data.

**Objectives**

This work package aims to characterize web content for what concerns audio and audiovisual data. As the use of multimedia channels is growing on the web, this work package provides an important dimension concerning what information is available missing from text-based sampling and characterization.

This work package has 4 tasks. Task 5.1 is concerned with specifying a representative corpus that will serve to train and evaluated statistical models for language identification. Task 5.2 concerns the annotation of this corpus with language and accent tags. Innovative methods will be explored to obtain the labelled data by incorporating speech technologies in the annotation process. Annotating the selected corpus is in and of itself a challenge, as few people are able to accurately tag more than a small number of languages. We propose a novel annotation procedure which will use an audio partitioner to chop the audio file into segments, which will then be clustered by speaker and language prior to presentation to humans for annotation. This type of task is well suited to annotation via crowd-sourcing, where segments can be presented to users until a consensus is found. VOCAPIA Research and CNRS will work together to develop this annotation framework. In coordination with Task 5.3, initial systems will be used to reduce the manual annotation load.

Prototype language identification systems will be developed in Task 5.3 and made accessible to the partners via a web-based service. Performance will be assessed on representative test data selected by the partners.

Task 5.4 is concerned with developing models for automatic content (topic) identification in audio data based on automatic speech-to-text transcription for the 8 covered languages. In the first year of the project, models will be developed for languages where STT systems already exist (French, Dutch, English, Spanish, German and Italian), taking into specificities of spoken language. Different methods will be explored to extract reliable content information from the automatic transcriptions which will always contains some recognition errors. Topic or content labels, selected in coordination with the other WP's, will be annotated in the automatic transcripts. Once STT systems are developed for the additional two SEMACORE languages (Catalan and Turkish), the most successful modeling techniques will be transferred to these languages. The automatic transcripts provided via the transcription service will also be used as text input to allow the developed opinion and sentiment analysis methods of WP6 to be applied to audio and audiovisual documents.

### *WP6 - Content analysis on multimedia data*

Summary:
This work package will process textual documents from a variety of sources (user PCs, smart phones, internet) and media (speech transcriptions from audio and video, web pages, documents) with the goal of extracting language-independent content analytics results suitable for the needs of the user cases. It will train models for the extraction of semantic information (language, document categories, disambiguated keywords, opinions, relational information) for multiple languages and text genres. It will

provide a near real-time workflow for the collection, semantic processing and evaluation of internet documents.

For the **crawling of web documents** in task T6.1, we will use an open source crawler (e.g. e.g. Apache Nudge, Bixo[30]) and use seed lists obtained from the use case partners the Common Crawl and other resources. Depending on the application we will perform focused crawling to prefer pages with the relevant topics [Olsten 2010]. For countries having too large a web domain for systematic crawling, some methods will be adapted (such as streaming sampling [Nesreen 2012]) to collect valid samples of web documents. We will employ the language identification component in FreeLing (UPC open-source library)[31] and the comparison of hash values for content sections to detect duplicates. Language identification operators are also provided for offline use in User Centered Measurement (WP4) devices.

All contents will be translated into a common interface format containing metadata (source, date, etc.), the content as a string / audio representation and the annotations (token, sentences, pos-tags, keywords, opinions, categories, etc.). Fraunhofer has a lot of experience with the CAS format employed by the Apache UIMA framework[32]. The final representation will be fixed in the system requirements.

For the different models (categories, opinion mining) different corpora of **training data** in T6.2 have to be provided. These have to be use-case specific, targeted to public issue texts and product-oriented areas. We will use manual annotation of text provided by the use case partners, which, for instance, can be entered via the brat rapid annotation tool[33]. In some cases we may need foreign language annotations which may be collected and validated by crowd-sourcing platforms[34]. Most of the training data will be generated by **machine translation**. We will use state-of-the-art open-source tools such as Moses[35], GIZA++[36], and SRLIM[37] to train the models. Training data will be obtained from Europarl[38] for European languages, and from UN corpus and NIST for Arabic. Available specific resources will be searched for Turkish and other languages not included in any standard data set used by the MT community. Translation dictionaries will be obtained from rule-based translation systems, and from bilingual dictionaries and ontologies such as WordNet or Wiktionary. We will also investigate if the translation of documents to a resource rich language and the extraction of information in that language is a accurate and efficient alternative.

For **keyword disambiguation** in T6.3 we will employ multilingual topic models based on parallel word-aligned documents generated by human and machine translators. We will investigate if the inclusion of entities (e.g. from Wikipedia) will improve disambiguation while still being not too time-consuming. For multilingual keyword search we will employ the possible translations of a keyword together with the corresponding topic distribution. Without additional effort this provides a monolingual search over synonyms.

For **text categorization** the training data provided by the use cases will be translated and will be annotated by multilingual topics. Then document classifiers will be trained, e.g. with Support Vector Machines (SVM). As an alternative we will evaluate multilingual topic models for direct classification. For named entity recognition we will also employ translated topic annotated training data. We will compare this to the novel multilingual approaches using Wikipedia labels in a semi-Markov CRF.

30 http://nutch.apache.org/ or http://openbixo.org/
31 http://nlp.lsi.upc.edu/freeling
32 Unstructured Information Management http://uima.apache.org/
33 http://brat.nlplab.org/
34 e.g. the Amazon Mechanical Turk, https://www.mturk.com
35 http://www.statmt.org/moses/
36 http://www.statmt.org/moses/giza/GIZA++.html
37 http://www.speech.sri.com/projects/srilm/
38 http://www.statmt.org/europarl/

As a baseline for **opinion mining** in T6.4 we will use conditional random field models trained on translated corpora. As an alternative we will implement multilingual deep learning models capturing the structure of a sentence. In addition the effect of parsetree-based models will be evaluated in one resource-rich language, where the sentences to be annotated will be translated to that language.

The key challenge for task T6.5 is to strike the right balance between evaluation using test collections, focused on components, and direct (interactive) evaluation involving user partners and their use-cases. T6.5 task also addresses aggregation of the raw entity mentions (of products and organisations) and their related opinions into a form that is better suited for presentation and use in WP7, taking the context of entity mentions into account (genre and content category, social context regarding who posted the mention as well as data context such as the related document in a twitter stream or forum site).
As a large proportion of web searches has been shown to be related to entities, and, entities and their attributes provide the anchors in interactive retrieval systems and their faceted search interfaces, a few scientific test collections involving entities have been created in recent years, including INEX 2007-2009 and TREC 2010-2011. TREC 2012 introduced a new track (Knowledge Base Accelleration, or KBA), to evaluate the filtering of a large news media stream on entity occurrences (close to envisioned use-cases in SEMACORE) [Frank2012].
However, SEMACORE goes beyond the scope of current entity centric search evaluations: the multi-lingual focus of the project, the importance of opinions, time, and regional interest, and the links in social networks, require a broader effort in evaluation. While specific components may be evaluated in context of the TREC KBA, the SEMACORE interests in multiple languages and geo-temporal relevance dimensions provide an excellent basis to propose a new lab at CLEF, Europe's TREC counter-part.
Where suitable public data and information needs would not be available to initiate building a new test collection (either directly from or inspired by our use-cases), we evaluate the remaining aspects using two strategies. First, we will evaluate our methods interactively, setting up the appropriate user studies in context of WP7's use cases. This method has the advantage that it most directly measures the contribution of SEMACORE to the actual end user's tasks. Second, we can use collected (non-public) data for predicting held-out data, in cross-validation, along the lines how social media search and annotation has been evaluated in [Clements2009]. The primary advantage here is to enable parameter tuning and testing technical aspects without requiring direct user involvement.
End users need support for **aggregation** of the raw data into a higher, more synthesized level than the annotated entity mentions identified stored in the repository as output of T6.3 and T6.4. Work package WP6 therefore also investigates how to show results in context of their use, both with respect to the origin of an entity mention (and its attributes) and its role in online conversation; the value and impact of a negative comment on a tweet for example, can only be interpreted correctly if we know who tweeted, whether they usually express negative information, if they have a complicated relation to a certain brand, how many people the tweet is expected to reach, what situation triggered the negative mention (e.g. a news document, or a forum site), etc. Communicating such aggregated information requires a careful design, making optimal use of state of the art web techniques. At a more fundamental level, we need to explore how to apply visualization techniques like timelines and maps over uncertain annotations, in such a way that the meaning of the underlying data can be conveyed correctly to end-users – clarifying also the uncertainty values inherent to content analysis results and their effect on the reported results.

Integration of the separate analysis modules in WP6 will be enabled by an advanced workflow management provided by WP3 that will orchestrate the main components and will utilize semantic standards to ensure the seamless integration of heterogeneous components and legacy systems. It will be backed by the stream engine and big data storage engine provided by WP3.
All models are tested separately and jointly on independent test data. In addition their resource consumption is determined.

***Applicative work package overview***

### *WP2 – Societal Issues*

Summary:
A major applicative focus of SEMACORE's characterization of the web concerns language diversity. In this work package, MAAYA defines the societal and linguistic indicators needed by policy makers. Intellectual property and privacy matters are also addressed here.

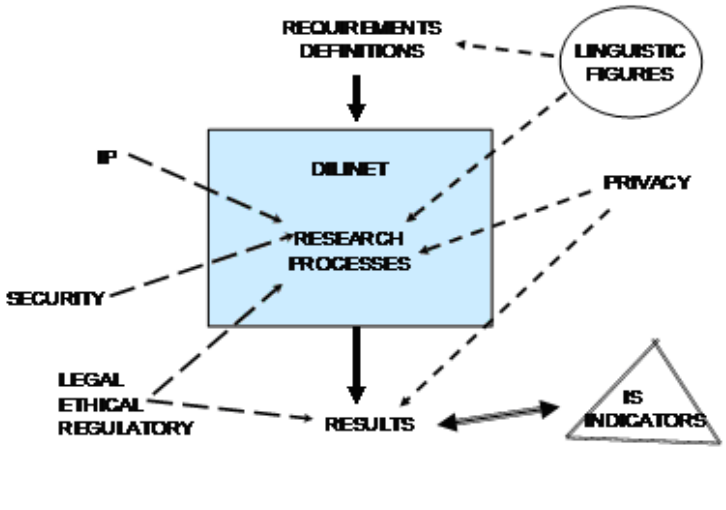


**Figure 8 - Focus of WP4 Societal Issues**

In this work package, all the side considerations that are required to strengthen the research and insert it into the appropriate societal contexts will be considered. The concept is of a comprehensive approach to enhance the integrity of the products of the developed research and relate them to various relevant societal realms (digital economy, linguistic policies, information society indicators…) in order to obtain impacts from the research beyond the theme of intelligent data management. The work package will deal with a series of matters which may affect the requirements for the research, the process of the research or the results itself.

Those matters could be classified as:

- context setting (linguistic figures),
- protection of all stakeholders interests (intellectual property, legal, ethical, regulatory, security and privacy).

#### *Intellectual properties issues*

This work-package will address different types of intellectual property and know-how issues that will arise during and result from the project implementation with the objective to maximize public domain decision when applicable and share the rights within the consortium. Intellectual property further covers intellectual property rights and know-how owned by the participants before the start of the project, and also intellectual property and know-how created outside of the project, during its duration, but connected to the project. Appropriate intellectual property protection will be considered, also taking into account possible copyright protection, patentability. It further covers any other kind of intellectual property protection related to software and more generally any kind of know-how that will be produced during the course of the project, as a result of the same or as part of any of its outputs. This includes know-how and any kind of intellectual property right developed in relation to a specific stage of the project or as an output of the same and will be protected through appropriate procedures and agreements established among the members of the Consortium. It may also be defined as a creative common, when appropriate. Management and protection of knowledge and intellectual property will be facilitated, within the project, by the close interaction with the task leader. From the early stage until the end of the project, the legal department of the partners will provide continuous assistance as to management of the knowledge produced and protection of intellectual property rights in any way arising or connected with the project.

#### *User Centered Measurement ethical issues*

This work-package will address any legal, regulatory or ethical issue raised by the research conducted in the project as well as by the products. Particular attention will be given to the User Centered Measurement Programs (WP4): the panels will be explicitly informed (through End User License Agreement/Terms of Use) about legal implications and ethical measures taken to ensure the protection of personal data and the statistical usage of the collected data within the framework of the project.
The design of the User Centered Measurement programs will be checked carefully so as not to capture or measure any unwanted information or identifiable data (such as IP numbers). The risk of the program being affected by specific viruses will be analyzed and appropriate protection to limit the risk will be put in place. The architecture of the user centered measurements foresees a client that is locally installed on one or more devices of the panellist, a central server that stores all the aggregated information, and a CRM-system that manages all the interactions with the panellists. The latter includes the sending of targeted survey questions to specific panellists via the client. There is thus two-way data exchange between the local client and the central server and the CRM-system. Task 2.3 will audit and address any privacy and security risks that are involved in the data exchange to and from the client. Four types of risks may be identified: the identification of the panellist based on the data that is being sent from the client; the interception of data during the exchange; the delivery of unwanted data to the client (including malevolent scripts that can be used to take over the client) and the exploitation of the security hole in the server, allowing a hacker to compromise the system. This WP will identify appropriate protection measures to be implemented, respectively for the anonymity of data (e.g., matching panellists based on a hash table, not on their IP address), for the encryption of data, and for the protection of the client and the server.

### *Linguistic diversity context data*

The gathering and compiling all required context data about linguistic matters produced outside the project (such as demographics and country policy assessment) will serve as a pool service to other work-packages requirements on linguistic matters and will provide the corresponding content of the web site to transform it into a clearinghouse on the subject of linguistic diversity on the digital world. This WP will also input WP7 (uses cases) with appropriate data and chronological series to allow assessment of the results produced by the research. A systematic study and documentation of the state of the art in the subject of linguistic diversity in the digital world will be conducted, considering both aspects relevant to public policies and business.

### *Information Society indicators*

The production of linguistic diversity indicators will be set in the broader context of the production of Information Society indicators which the project will interface in this work package via MAAYA through institutional parties (ITU, UNESCO). It will in particular address, within the follow-up of the World Summit of Information Society (WSIS), the WSIS target 9, which is to “encourage the development of content and put in place technical conditions in order to facilitate the presence and use of all world languages on the Internet”, and WSIS action line C8 (cultural and linguistic diversity). The topic of indicators on linguistic diversity is also addressed by the Task Group on Measuring the WSIS targets, which is part of the Partnership on Measuring ICT for Development, a multi-stakeholder initiative to improve the availability and quality of ICT data and indicators. This task group will allow feedback into the working group on multilingualism of the Broadband Commission for Digital Development (http://www.broadbandcommission.org).

### *WP7 – Use Cases*

Summary:
The objective of this work package is the development of a set of applications that will use SEMACORE data and evaluation methods in real life situations. Those cases include the identification of actual trends and the change of language, content, and opinions for international public agencies and companies as well as the measurement of linguistic diversity and its correlation with other factors in various contexts (international websites, digital libraries, countries webspaces, a selection of blogs).

Whenever applicable the data will be produced in forms of graphs and diagrams. International public agencies, companies and policy making organizations can use the analysis and results out of this work package to develop, promote and protect their content, products, services or policies as well as identifying new trends for innovation or policy changes. Results and analyses within this work package will provide volume metrics, opinion driving indicators, and statistics of opinion/sentiment (good, bad, neutral) on specific concepts and context information (interrelation between language and/or opinion/sentiment information). Furthermore WP7 will provide the project with the means to assess the validity of the data produced from proven data and through the evaluation of the cases offers a forecast model for future impacts.

### *Use Case for linguistic diversity*

Four different contexts will be scrutinized with the criteria of linguistic diversity measurement:

1. The web sites of European Union and United Nations, to measure proportions of languages usage.
2. The main global digital libraries, to measure the respective proportions of documents in different languages.
3. The full National Web Space of Four European countries which will be fully characterized.
4. A selection of European blogs which will allow checking citizen's opinions on languages.

### *Use Case for international public agencies*

European Commission DG SANCO: SEMACORE will collect the data for a special topic (e.g. Health or Hydration) at different time points in different countries. Audio documents will be prepared by speech-to-text translation. Subsequently a categorization of the text with respect to 150 content classes is performed. The application will detect a number of key phrases (e.g. product names, issue names, EU or national initiatives) together with the content category that will be prepared for the selected topic as well as relevant named entities (e.g drugs). Furthermore for specific keywords the opinions expressed on these keywords (positive, negative, neutral) will be extracted to get an indication of the positioning of key phrases (e.g. product names, issue name or EU or national initiatives). Opinions and wishes of citizens on the selected special topic related to the public agency are available to the decision makers to be taken into consideration. In the other direction the agency could publish questions to specific forums to discuss alternatives. The information will be transferred in dedicated graphs to provide deep insights into the topic and a flexible evaluation GUI will be provided.. One solution will be association maps that display the information on quantitative as well as qualitative level around the special topic.

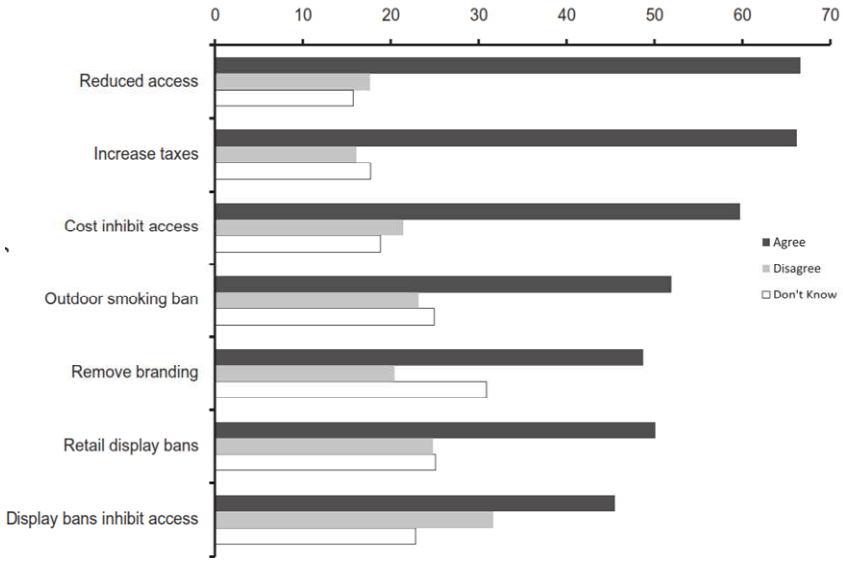


**Figure 9 – Example: Opinions on different aspects of smoking**

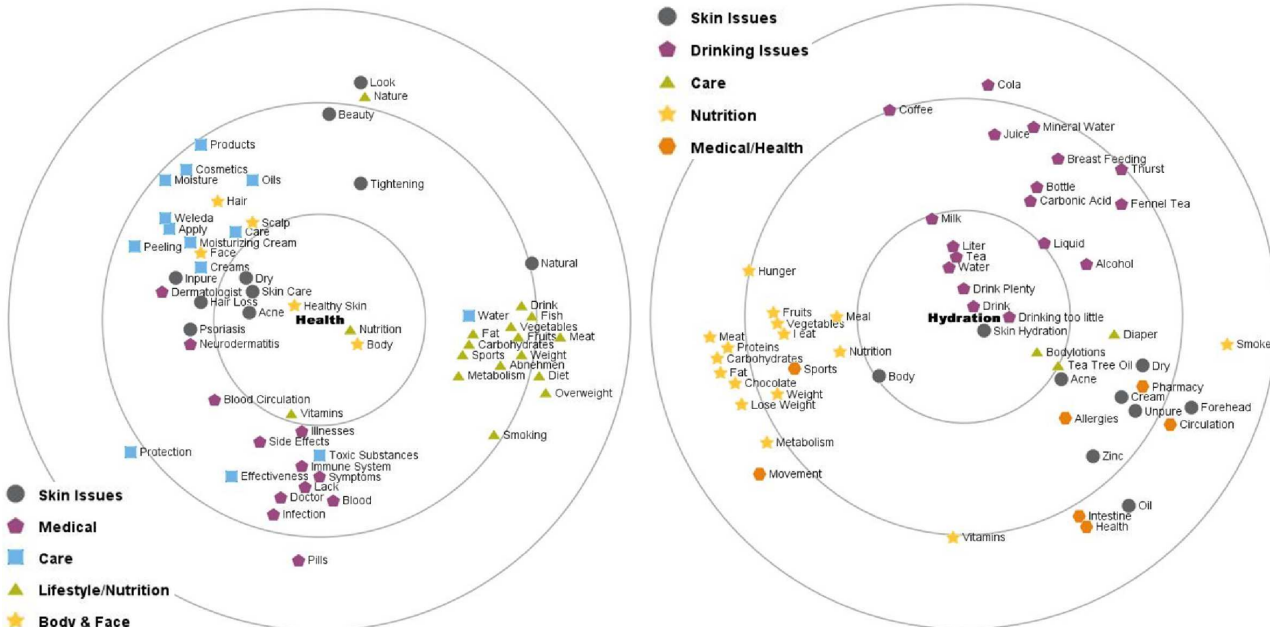


**Figure 10 - Two example association maps on health (left) and hydration (right)**

### *Use Case for international companies*

International companies have observed that the consumer is the new authority. Due to social media and other digital information and communication services, the consumer is well informed and has plenty of opportunities to share their opinions on services and products. An international company (e.g. fast moving consumer goods) will receive deep information regarding customer satisfaction and problem categories, sentiment analysis (positive, neutral, negative), sentiment variance between different blogs, competitors positioning, based on web data collected and analyzed by SEMACORE. Business will be enabled to identify trends or challenges issues when they are starting, and incorporate solutions and updates in their brands early in the process.

Social media is transforming the way that consumers across the globe make purchase decisions. Consumers around the world are using social media to learn about other consumers' experiences, to find more information about brands, products and services, and to find deals and purchase incentives. The days when companies could tightly control brand messaging and progress consumers along a linear purchase funnel have long ended. Social media has fundamentally changed the consumer decision journey. Consumer decisions and behaviours are increasingly driven by the opinions, tastes and preferences of an exponentially larger, global pool of friends, peers and influencers. This use case among others will investigate the impact of real-time, authentic consumer expression in social media on the development of strong brands, the emergence of passionate and engaged brand communities, and the ultimate effect on sales.

Nielsen supports many international companies in market analysis and product lifecycle management. If marketing takes place on a European scale it is of utmost importance to receive timely and reliable feedback of customers on brands and company reputation and to capture rumors and trends discussed by customers. In this use case specific products and brands of a single company will be addressed. The number of mentions of the products in different types of internet genres, languages, regions, and content categories will be determined for text and speech. In addition opinions on products will be collected similar to the figure below. Here the opinion with respect to aspects grouped in broader categories have been determined and are presented to the users in a dynamic graph. Moreover comparisons of a product with competitor products will be analysed. Nielsen's reference list for Social Media analysis is including, for example, Procter & Gamble, Disney, Honda, Western Union, 3M, and BBVA.

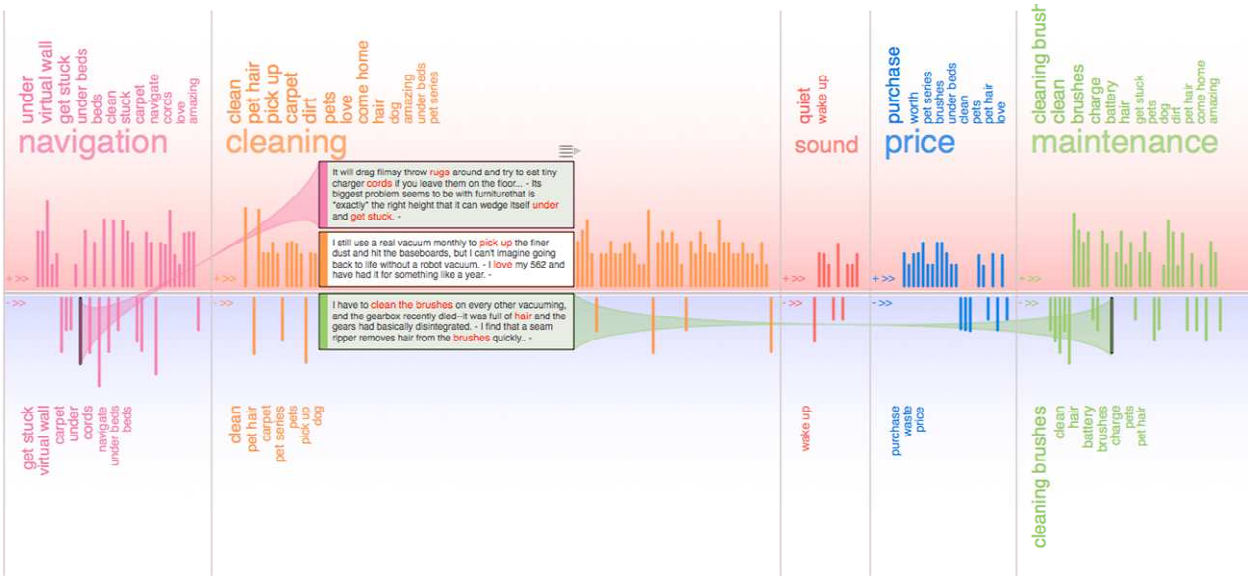


**Figure 11 - Example showing opinions from consumer reviews grouped by major product features (colour coded) that are discussed most frequently in the reviews of a vacuum cleaner robot. Keywords for each feature-sentiment group are shown in the positive (upper half) and negative (lower half) region. The height of the lines is proportional to the number of mentions of each feature. A single snippet is expanded, so that it shows other snippets from the same review [Alper 2011].**

### *WP8 - Dissemination & Exploitation*

Summary:
The purpose of this work package is the dissemination and exploitation of the progress and results of the work achieved in the project. Detailed dissemination, exploitation and use plans will be compiled to establish the various stages in the process. The dissemination will be realized both directly in a comprehensive web site reflecting the progress and outcomes of the research and indirectly through appropriate publications targeting the research and the business communities as well as public at large. Particular attention is given to the economic exploitation, the provision of a standardized workflow for including new languages in the joint analysis and an effort to warrant the sustainability of the language indicators production. Two workshops will be organized with relevant stakeholders, one with a policy focus, the other with a science and business focus; both will serve to strengthen respective exploitation strategies.

**Dissemination:** The expected results from SEMACORE will be disseminated in a number of ways, leveraging on the partners gathered in the project and on the international organisations affiliated to the project. A public SEMACORE website will contain information on (the progress of) the project, published reports, papers made available to the research community. Planned initiatives include publication in scientific journals, presentations at conferences, networking. Two workshops will be organized on scientific and use case results. Consortium members will also participate in clustering activities, in order to exchange ideas with projects active in the same area; appropriate brochures, leaflets, videos and presentation supports will be prepared to communicate on the project progress.

**Exploitation:** This includes the planning and start of the technological exploitation of the scientific innovations resulting from the work accomplished in SEMACORE. Exploitation of the final results of the project will happen via all partners, industrial as well as institutional. A proof of concept will demonstrate the concept of the developed technologies. Plans will be made with regards to marketing and potential implementation and deployment. As for the potential of the project for international organizations which invested via MAAYA in the project, a specific effort will be accomplished to allow

the generalization to other languages and countries and the systematization of the data production. At this effect a sustainability roadmap will be developed and a foresight exercise will be realized.

### B 1.3.5 List of Work packages

| WP N° | Work package title | Type of activity | Lead partic no. | Lead partic. short name | Person-months | Start month | End month |
|---|---|---|---|---|---|---|---|
| WP 1 | Project Management | MGT | 1 | ERCIM | 14 | M1 | M36 |
| WP 2 | Societal Issues | RTD | 2 | MAAYA | 24 | M1 | M36 |
| WP 3 | System Development & Integration | RTD | 7 | FRAUNHOFER | 49.5 | M1 | M36 |
| WP 4 | User Centric Measurement | RTD | 4 | DIALOGIC | 39 | M1 | M36 |
| WP 5 | Analysis of spoken language in multimedia data | RTD | 5 | CNRS | 32.5 | M1 | M36 |
| WP 6 | Content analysis on multimedia data | RTD | 3 | UPC | 101 | M1 | M36 |
| WP 7 | Use Cases | RTD | 10 | NIELSEN | 66.5 | M1 | M36 |
| WP 8 | Dissemination & Exploitation | OTH | 2 | MAAYA | 27.6 | M1 | M36 |
| | TOTAL | | | | 354.1 | | |

**Figure 12 - Work package detailed list**

### B 1.3.6 List of deliverables

| Del. N° | Deliverable name | WP N° | Nature | Dissemi -nation level | Delivery date |
|---|---|---|---|---|---|
| D1.1 | Collaborative working environment documentation | 1 | R | PU | M3 |
| D1.2 | First bi-annual activity report | 1 | R | PU | M7 |
| D1.3 | Periodic report Year 1 | 1 | R | PU | M12 |
| D1.4 | Second bi-annual activity report | 1 | R | PU | M19 |
| D1.5 | Periodic report Year 2 | 1 | R | PU | M24 |
| D1.6 | Third bi-annual activity report | 1 | R | PU | M31 |
| D1.7 | Periodic report Year 3 | 1 | R | PU | M36 |
| D1.8 | Final project reports | 1 | R | PU | M36 |
| D2.1 | Document of understanding for IP matters | 2 | R | PU | M6 |
| D2.2 | Document of understanding for legal matters | 2 | R | PU | M6 |
| D2.3 | Document of understanding for ethical matters | 2 | R | PU | M6 |
| D2.4 | Linguistic figures version 1 (plus inputs for indicators specs) | 2 | R | PU | M6 |
| D2.5 | Document of recommendations for security matters in WP4 | 2 | R | PU | M12 |
| D2.6 | Stress test reports of security for WP4 | 2 | R | PU | M18 |
| D2.7 | Linguistic figures version 2 | 2 | R | PU | M24 |
| D2.8 | Strategy document for information society indicators | 2 | R | PU | M36 |
| D2.9 | Linguistic figures final version | 2 | R | PU | M36 |
| D3.1 | Terminology Definition Report | 3 | R | PU | M2 |
| D3.2 | System Requirements Report | 3 | R | PU | M04 |
| D3.3 | First Stable Prototype | 3 | PR | PU | M12 |
| D3.4 | Updated System Requirements and Architecture | 3 | PR | PU | M18 |
| D3.5 | Second Stable Prototype | 3 | PR | PU | M18 |
| D3.6 | Final System Requirements and Architecture | 3 | R | PU | M30 |
| D3.7 | Third Prototype | 3 | RP | PU | M36 |
| D4.1 | Report on scope of the project in terms of number and types of operating systems supported | 4 | R | PU | M4 |
| D4.2 | First stable version of protocol | 4 | R | PU | M4 |
| D4.3 | Mailing list of panel (initial set of respondents) | 4 | R | RE | M20 |
| D4.4 | Audit report on privacy and security aspects | 4 | R | PU | M19 |
| D4.5 | User-centric measurement clients (both desk top and mobile based), including survey module and back end interface | 4 | P | RE | M36 |
| D4.6 | Final report based on the analysis of the user-centric measurements | 4 | R | PU | M36 |
| D5.1 | Corpus Specification | 5 | R | PU | M3 |
| D5.2 | Report on Annotated Corpus | 5 | R | PU | M12 |
| D5.3 | Report on Language Identification in audio | 5 | R | PU | M25 |
| D5.4 | Report on Topic Detection in audio | 5 | R | PU | M35 |
| D6.1 | Specification of APIs and integration design | 6 | R | PU | M6 |
| D6.2 | Report on Crawler prototype | 6 | R | PU | M12 |
| D6.3 | Report on Machine Translation prototype | 6 | R | PU | M12 |
| D6.4 | Report on Linguistic enricher prototype | 6 | R | PU | M18 |

| | | | | | |
|---|---|---|---|---|---|
| D6.5 | Report on Disambiguation and Classification prototype | 6 | R | PU | M12 |
| D6.6 | Report on Opinion Mining Prototype | 6 | R | PU | M18 |
| D6.7 | Report on Aggregation and evaluation prototype | 6 | R | PU | M18 |
| D6.8 | Content Analysis prototype | 6 | P | PU | M30 |
| D6.9 | Report on the Workflow for the Extension of the Analysis by another Language | 6 | R | PU | M36 |
| D7.1 | Initial progress report on SEMACORE Use Cases covering all sources described within WP7 | 7 | R | PU | M12 |
| D7.2 | Intermediary progress report on SEMACORE Use Cases | 7 | R | PU | M24 |
| D7.3 | Final progress report on SEMACORE Use Cases | 7 | R | PU | M36 |
| D7.4 | Report on Funredes methods measurements and cross-checking and validation of project produced data | 7 | R | PU | M36 |
| D7.5 | Strategy paper on forecast impact | 7 | O | PU | M36 |
| D7.6 | Report on Use Cases evaluation | 7 | R | PU | M36 |
| D8.1 | Initial project web site | 8 | O | PU | M3 |
| D8.2 | Dissemination and use plan V1 | 8 | O | PU | M6 |
| D8.3 | Dissemination and use plan V2 | 8 | O | PU | M18 |
| D8.4 | Strategy document for research roadmap | 8 | R | PU | M30 |
| D8.5 | Strategy document for indicator sustainability | 8 | R | PU | M30 |
| D8.6 | Final dissemination and exploitation report | 8 | O | PU | M36 |

### B 1.3.7 List of milestones

| Milestone number | Milestone name | WP(s) involved | Expected date | Means of verification |
|---|---|---|---|---|
| M2.1 | Partner's meeting for consensus on legal, ethical, privacy and security matters | 2 | M12 | Meeting evaluation |
| M2.2 | Partner's meeting for finalization on specification for indicators | 2 | M12 | Meeting evaluation |
| M2.3 | WP4 security under control | 2 | M18 | Stress test report |
| M2.4 | Partner's meeting for strategy decisions for languages matters and information society | 2 | M36 | Meeting evaluation |
| M3.1 | First Runnable Prototype | 3 | M6 | Automatic software tests and metrics and manual inspection |
| M3.2 | Stable Prototype 1 | 3 | M12 | Automatic software tests and metrics and manual inspection |
| M3.3 | Stable Prototype 2 | 3 | M24 | Automatic software tests and metrics and manual inspection |
| M3.4 | Stable Prototype 3 - System Version 1 | 3 | M36 | Automatic software tests and metrics and manual inspection |
| M4.1 | Baseline protocol ready | 4 | M5 | Protocol has been thoroughly tested internally and tested successfully on an external panel |
| M4.2 | Baseline panel ready | 4 | M12 | A panel has been composed that meets the critical statistical size in |

| | | | | |
|---|---|---|---|---|
| | | | | each country involved |
| M4.3 | Beta version 1.0 of client ready | 4 | M18 | First version of the client available, technically tested (bug free) |
| M4.4 | Beta version 1.1 of client ready | 4 | M19 | Second version of the client available, with survey module integrated, technically tested (bug free) |
| M4.5 | Panel active for 6 months (halfway full duration) | 4 | M25 | Client installed at devices from 10,000 users in 10 countries, with a consistent 90% of active users |
| M4.6 | Debriefing of panel (end of duration, after 12 months) | 4 | M31 | All 10,000 users informed about the completion of the measurement. |
| M5.1 | Baseline web-based LID service available | 5 | M6 | Automatic software tests and metrics and manual inspection |
| M5.2 | Web-based STT service available (6 languages) | 5 | M6 | Automatic software tests and metrics and manual inspection |
| M5.3 | Baseline web-based topic detection service available (X languages) | 5 | M18 | Automatic software tests and metrics and manual inspection |
| M5.4 | Final web-based LID service available | 5 | M24 | Automatic software tests and metrics and manual inspection |
| M5.5 | Final web-based topic detection service available (XX languages) | 5 | M30 | Automatic software tests and metrics and manual inspection |
| M6.1 | Baseline web-based crawler service available | 6 | M12 | Tests by independent test sets with known result. |
| M6.2 | Baseline web-based MT service available | 6 | M12 | Tests by independent test sets with known result. |
| M6.3 | Baseline web-based linguistic enricher service available | 6 | M18 | Tests by independent test sets with known result. |
| M6.4 | Baseline web-based classification service available | 6 | M12 | Tests by independent test sets with known result. |
| M6.5 | Baseline web-based opinion mining service available | 6 | M18 | Tests by independent test sets with known result. |
| M6.6 | Baseline web-based content analysis tools available | 6 | M18 | Tests by independent test sets with known result. |
| M6.7 | Final web-based crawler service available | 6 | M24 | Tests by independent test sets with known result. |
| M6.8 | Final web-based MT service available | 6 | M24 | Tests by independent test sets with known result. |
| M6.9 | Final web-based linguistic enricher service available | 6 | M28 | Tests by independent test sets with known result. |
| M6.10 | Final web-based classification service available | 6 | M30 | Tests by independent test sets with known result. |
| M6.11 | Final web-based opinion mining service available | 6 | M30 | Tests by independent test sets with known result. |
| M7.1 | Use Cases planning | 7 | M12 | Report evaluated by EAB |
| M7.2 | Software programming and test of T7.4 | 7 | M18 | Test success |
| M7.3 | Uses cases launch | 7 | M18 | Partner meeting |
| M7.4 | Uses cases completion | 7 | M30 | Reports |
| M7.5 | Uses case evaluation and impact forecast | 7 | M36 | Reports |

| | | | | |
|---|---|---|---|---|
| M8.1 | Consolidated version of web site | 8 | M12 | Online review |
| M8.2 | Consolidated plan for all dissemination and exploitation activities | 8 | M18 | Partner meeting |
| M8.3 | Review of plan for all dissemination and exploitation activities | 8 | M24 | Partner meeting |
| M8.4 | Workshop for strategies on business | 8 | M30 | Web site |
| M8.5 | Workshop for strategies on policies | 8 | M30 | Web site |
| M8.6 | Evaluation of all dissemination and exploitation activities | 8 | M36 | Report |

### B 1.3.8 Tabular description of work packages

| **Work package number** | **1** | | **Start - End:** | | M1 – M36 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Work package title** | **Project Management** | | | | | | | | | |
| **Activity type**[39] | MGT | | | | | | | | | |
| **Participant number** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| **Participant short name** | **ERCIM** | MAAYA | UPC | DIALOGIC | CNRS | KYOS | FHG | CWI | VOCAPIA | NIELSEN |
| **Person-months per participant** | **14** | | | | | | | | | |

**Objectives**

The objective of this work package is to ensure an efficient administration and co-ordination of all project activities, focused to the objectives of the project.

**Description of work**

**Task 1.1 – Co-ordination of activities (M1-M36, 3 PM)**

*Task Leader: ERCIM*

*Participants:*

This task is to plan and monitor the project activities, in close collaboration with the Scientific Coordinator, ensuring an effective coordination, detecting early possible deviations, and appropriately addressing these. It includes the organisation and participation in (physical or virtual) project meetings. Specific measurable indicators will be defined in order to monitor overall project progresses with respect to the global project objectives. This task includes monitoring of non-technical risks and taking the necessary actions for risk mitigation.

In order to run the project in a professional manner, the procedures that will be put in place will be described in "Management Notes" that will be available in the Collaborative web environment.

This task consists also of preparing the Consortium to official Project review meetings, ensuring timely provision of information, good quality and professional presentations and rehearsal sessions.

The coordinator will chair the General Assembly.

**Task 1.2 – Resource management (M1-M36, 3 PM)**

*Task Leader: ERCIM*

*Participants:*

This task is to manage the project resources and specifically the project budget. It consists of administrating the Community financial contribution, monitoring the use of resources by each partner, and validating transfer of budget between activities and beneficiaries.

**Task 1.3 – Project reporting (M5-M36, 3 PM)**

*Task Leader: ERCIM*

*Participants:*

[39] Please indicate one activity per work package:
RTD = Research and technological development; DEM = Demonstration; MGT = Management of the consortium; OTHER = Other specific activities if applicable to this call, including any activities to prepare for the dissemination and/or exploitation of project results and coordination activities.

This task is to produce the set of deliverables to the European Commission on a regular basis. The task consists of, on one end to gather the contribution by all partners, and on the other end to produce the periodic and final project reports.
The periodic reports shall comprise:

- An overview, including a publishable summary of the progress of work towards the objectives of the project, including achievements, milestones and deliverables). If there were deviations from the original work plan, this report includes their description and motivation.
- An explanation of the use of the resources.
- A Financial Statement from each partner together with a summary financial report consolidating the claimed Community contribution of all the project partners in an aggregate form.

This final report shall comprise:

- A final publishable summary report covering results, conclusions and socio-economic impact of the project.
- A report covering the wider societal implications of the project, including gender equality actions, ethical issues, efforts to involve other actors and to spread awareness, as well as the plan for the use and dissemination of foreground.

All reports and deliverables will be published in English.

**Task 1.4 – Deliverable quality assurance (M1-M36, 3 PM)**
*Task Leader: ERCIM*
*Participants:*
In collaboration with the Project Co-ordinator, the task leader will organise and drive a quality assurance process for all project deliverables. The process will be as follows:

- Intended table of content of the deliverable is available 1 month after the corresponding task kick-off date.
- Draft deliverable versions are available four weeks before planned delivery date.
- Appointment of two internal reviewers from different organisations than the deliverable author(s).
- Reviewers carry out their reviews and produce a written (email) report within two weeks of reception of the draft deliverable.
- Author(s) take into account the reviewers comments and produce the final version of the deliverable.

This procedure will be documented as a “Management Note”. In all project management meetings, ERCIM will report on the quality assurance process to make sure that the group improves its performance.

**Task 1.5 –Collaborative tools (M1-M36, 2 PM)**
*Task Leader: ERCIM*
*Participants:*
In order to support an efficient collaboration between the project partners, we will set up and manage an appropriate project IT infrastructure, that will consist of 5 components:

- Mailing lists,
- Wikis,
- Collaborative web environment (BSCW or equivalent),
- Source code revision management (SVN),
- Issue Tracker (TRAC).

**Deliverables** (brief description) and month of delivery

| Number [40] | Description | WP | Nature [41] | Diss level [42] | Month Due |
|---|---|---|---|---|---|
| D1.1 | Collaborative working environment documentation | 1 | R | PU | M3 |
| D1.2 | First bi-annual activity report | 1 | R | PU | M7 |
| D1.3 | Periodic report Year 1 | 1 | R | PU | M12 |
| D1.4 | Second bi-annual activity report | 1 | R | PU | M19 |
| D1.5 | Periodic report Year 2 | 1 | R | PU | M24 |
| D1.6 | Third bi-annual activity report | 1 | R | PU | M31 |
| D1.7 | Periodic report Year 3 | 1 | R | PU | M36 |
| D1.8 | Final project reports | 1 | R | PU | M36 |

| **Milestones** | | | | |
|---|---|---|---|---|
| Number | Short description | WP | Month | Means of verification [43] |
| | N/A | | | |

Note: footnotes commenting the WP description table headings have been left visible on WP1 only.

[40] Deliverable numbers in order of delivery dates.

[41] Please indicate the nature of the deliverable using one of the following codes:

**R** = Report, **P** = Prototype, **D** = Demonstrator, **O** = Other

[42] Please indicate the dissemination level using one of the following codes:

**PU** = Public

**PP** = Restricted to other programme participants (including the Commission Services).

**RE** = Restricted to a group specified by the consortium (including the Commission Services).

**CO** = Confidential, only for members of the consortium (including the Commission Services).

[43] Show how you will confirm that the milestone has been attained. Refer to indicators if appropriate. For example: a laboratory prototype completed and running flawlessly; software released and validated by a user group; field survey complete and data quality validated.

| **Work package number** | **2** | | **Start - End:** | | M1 – M36 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Work package title** | **Societal Issues** | | | | | | | | | |
| **Activity type** | RTD | | | | | | | | | |
| **Participant number** | 1 | **2** | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| **Participant short name** | ERCIM | **MAAYA** | UPC | DIALOGIC | CNRS | KYOS | FHG | CWI | VOCAPIA | NIELSEN |
| **Person-months per participant** | | **15** | | 1 | | 8 | | | | |

**Objectives**

This work package gathers the focuses on all the side considerations that are required to guarantee the quality and validity of the research and their insertion onto the appropriate societal contexts. The side considerations are concerned with ethical and legal matters as well as those related to Intellectual Property, security, and privacy. As for the appropriate societal contexts, one is the linguistic diversity indicators in the digital world which requires to be contrasted with general data about linguistic diversity, and the other ones is the information society complex of indicators of which it is part and which paradigm shall be positively impacted by the progress made in the project. To allow posterior follow-up and production of indicators by MAAYA and associated international bodies, the precise description of the process, method, algorithms and required software to add a country or a language will be provided and whenever it is possible Open Source would be the preferred choice for software.

***Tangible outcomes and measures of progress and success***

The tangible outcomes include:

- a consortium agreement for intellectual property matters which will be evaluated by the External Advisory Board;
- a guidelines document on legal and ethical matters which will be evaluated by the External Advisory Board;
- a reference document for the security and privacy matters which will serve as input for WP4 and will be evaluated by the External Advisory Board;
- a report of the result of stress testing of security and privacy for the software provided in WP4 which will be evaluated by the External Advisory Board;
- a compilation of data and figures about linguistic diversity which will be actualized each year within the time frame of the project and organized in a section of the web site and which will be evaluated by the External Advisory Board;
- a report of the interaction of the project with the WSIS process and especially the task group in measuring WSIS targets which will be evaluated by the External Advisory Board.

**Description of work**

**Task 2.1** Intellectual property matters **(M1-M36, 2 PM)**

*Task Leader: KYOS*

*Participants:*

This task will address the formulation of the guidelines for protecting the Intellectual Property Rights (IPR) of consortium partners and will be part of the consortium agreement. They will consider different options to share the IPR across partners and external parties through appropriate license agreements.

**Task 2.2** Legal, ethical and regulatory matters **(M1-M36, 5 PM)**

*Task Leader: KYOS*

*Participants: MAAYA*
This task will address any legal, regulatory or ethical issue raised by the research conducted in the project in coordination with all partners. It will also analyze and describe the technical requirements on how data will be collected, and why and how it will be used in compliance with various EU and UN international normative documents as human dignity, integrity of the person, the right to privacy, etc. Particular attention will be given to the User Centered Measurement Programs (WP4) about legal implications and ethical measures taken to ensure the protection of personal data and the statistical usage of the collected data within the framework of the project.

**Task 2.3** Privacy protection and security technologies **(M1-M36, 7 PM)**
*Task Leader: KYOS*
*Participants: DIALOGIC, MAAYA*
This task will address privacy and security issues that are involved in the client environment, on the server side and in the data exchange. This task will identify appropriate protection measures to be implemented for the anonymity of data, the encryption of data, the security of communications and the protection of the client and the server environments. The User Centered Measurement programs will be designed so as not to capture or measure any unwanted information or identifiable data.

**Task 2.4** Linguistic diversity & information society indicators **(M1-M36, 10 PM)**
*Task Leader: MAAYA*
*Participants:*
This task consists in two parts.
The first part is the gathering and compiling all required context data about linguistic matters produced outside the project (such as demographics and country policy assessment) to feed other WPs which require such data.
In the second part the production of linguistic diversity indicators will be set in the broader context of the production of Information Society indicators which the project will interface via MAAYA through institutional parties and corresponding bodies.

**Deliverables** (brief description) and month of delivery

| Number | Description | WP | Nature | Diss level | Month Due |
|---|---|---|---|---|---|
| D2.1 | Document of understanding for IP matters | 2 | R | PU | M6 |
| D2.2 | Document of understanding for legal matters | 2 | R | PU | M6 |
| D2.3 | Document of understanding for ethical matters | 2 | R | PU | M6 |
| D2.4 | Linguistic figures version 1 (plus inputs for indicators specs) | 2 | R | PU | M6 |
| D2.5 | Document of recommendations for security matters in WP4 | 2 | R | PU | M12 |
| D2.6 | Stress test reports of security for WP4 | 2 | R | PU | M18 |
| D2.7 | Linguistic figures version 2 | 2 | R | PU | M24 |
| D2.8 | Strategy document for information society indicators | 2 | R | PU | M36 |
| D2.9 | Linguistic figures final version | 2 | R | PU | M36 |

**Milestones**

| Number | Short description | WP | Month | Means of verification |
|---|---|---|---|---|
| M2.1 | Partner's meeting for consensus on legal, ethical, privacy and security matters | 2 | M12 | Meeting evaluation |
| M2.2 | Partner's meeting for finalization on specification for indicators | 2 | M12 | Meeting evaluation |
| M2.3 | WP4 security under control | 2 | M18 | Stress test report |
| M2.4 | Partner's meeting for strategy decisions for languages matters and information society | 2 | M36 | Meeting evaluation |

| **Work package number** | **3** | | **Start - End:** | | M1 – M36 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Work package title** | **SYSTEM DEVELOPMENT AND INTEGRATION** | | | | | | | | | |
| **Activity type** | RTD | | | | | | | | | |
| **Participant number** | 1 | 2 | 3 | 4 | 5 | 6 | **7** | 8 | 9 | 10 |
| **Participant short name** | ERCIM | MAAYA | UPC | DIALOGIC | CNRS | KYOS | **FHG** | CWI | VOCAPIA | NIELSEN |
| **Person-months per participant** | | 2.5 | 2 | 2 | 3 | | **34.5** | 2 | 1.5 | 2 |

**Objectives**

An operational infrastructure according to the Big-Data lambda architecture will provide a framework for technical experiments, for evaluation and enhancements. Appropriate coordination workflows of the batch layer and the speed layer will be explored. Training workflows for semantic models will usually be executed in the batch layer. The speed layer applies the workflows of trained semantic analysis methods to incoming data in real-time. Development operates with agile sprint and Kanban methods using a collaborative workspace and an issue tracking system.

***Tangible outcomes and measures of progress and success***

The tangible outcomes are three system releases at a rate of 12 months and available at the interim and final reviews. The requirements of these releases are documented in System Requirements and Architecture Reports. Manual inspection of software modules as well automatic software tests and metrics will be employed to generate an up-to-date assessment of functional and non-functional adequacy as well as the software quality of the different software modules provided by the different work packages.

- A first runnable prototype after six months that makes the proposed integration facilities tangible.
- Three stable system releases available at the annual reports to the commission, including System Requirements and Architecture Reports.
- The rationale of the WP4, WP5, WP6, and WP7 components integration into the architecture.
- Module and system tests
- Report on metrics of static and dynamic characteristics of the system

**Description of work**

**Task 3.1** Scientific coordination **(M1-M36, 6 PM)**
*Task Leader: FRAUNHOFER*
*Participants:*
This task is to carry out the planning, management and monitoring of project-wide research and technological development activities, including the coordination of scientific and technical work between work packages and use cases. Special attention will be given to monitoring technical risks and taking the necessary actions for risk mitigation. Day-to-day technical management of individual work packages is the responsibility of the WP leaders.

**Task 3.2** Terminology Definition and System Requirements **(M1-M36, 10.5 PM)**
*Task Leader: FRAUNHOFER*
*Participants: MAAYA, UPC , DIALOGIC, CNRS, CWI, VOCAPIA, NIELSEN*

The participants of the SEMACORE project have a very diverse professional background. To allow an unambiguous communication this task will establish an agreed set of common concepts which will be defined in a multilingual glossary.
Using this terminology the partners will analyse the functional and non-functional systems requirements arising from WP2 to WP7. The functional requirements describe the SEMACORE capabilities for solving a functional, application-specific problem. The non-functional requirements are not of a functional nature, but contribute decisively to the applicability of the system (e.g. quality, safety, or performance requirements). All partners discuss and specify the requirements and derive the requirements for the tools, services and the overall system.

**Task 3.3** System architecture **(M1-M36, 14 PM)**
*Task Leader: FRAUNHOFER*
*Participants: UPC, DIALOGIC, CNRS, CWI, VOCAPIA, NIELSEN*
The main objective of this task is to define a detailed architecture of the SEMACORE system. The system architecture will identify the components and services implemented by the project. The system architecture also specifies the main selected technologies, protocols, and middleware that are used in the project.

**Task 3.4** Project infrastructure, processes and tools **(M1-M36, 12 PM)**
*Task Leader: FRAUNHOFER*
*Participants:*
This task will specify and implement all infrastructure elements of SEMACORE. For coordination and documentation purposes these are the collaboration workspace and the issue tracking system. For product integration and operation these are interfaces, components and services allowing plugging and interoperation of external (legacy systems) and internal systems. The SEMACORE components developed in work packages 4-7 have to conform to the services, interfaces, and processes defined in this work package.

**Task 3.5** Prototype development cycle **(M1-M36, 7 PM)**
*Task Leader: FRAUNHOFER*
*Participants: MAAYA*
For the series prototypes provided by SEMACORE which are increasingly comprehensive, this task will monitor the compliance of modules to the standards defined in Tasks 3.1-3.4. The task will also check the adequacy of software tests and optimizations. Finally this task will update requirements arising during the course of the project in a concerted way and adapt the project terminology, infrastructure, processes, and tools accordingly.

**Deliverables** (brief description) and month of delivery

| Number | Description | WP | Nature | Diss level | Month Due |
|---|---|---|---|---|---|
| D3.1 | Terminology Definition Report | 3 | R | PU | M02 |
| D3.2 | System Requirements Report | 3 | R | PU | M04 |
| D3.3 | First Stable Prototype | 3 | PR | PU | M12 |
| D3.4 | Updated System Requirements and Architecture | 3 | PR | PU | M18 |
| D3.5 | Second Stable Prototype | 3 | PR | PU | M18 |
| D3.6 | Final System Requirements and Architecture | 3 | R | PU | M30 |
| D3.7 | Third Prototype | 3 | RP | PU | M36 |

**Milestones**

| Number | Short description | WP | Month | Means of verification |
|---|---|---|---|---|
| M3.1 | First Runnable Prototype | 3 | M06 | Automatic software tests and metrics and manual inspection |
| M3.2 | Stable Prototype 1 | 3 | M12 | Automatic software tests and metrics and manual inspection |
| M3.3 | Stable Prototype 2 | 3 | M24 | Automatic software tests and metrics and manual inspection |
| M3.4 | Stable Prototype 3 - System Version 1 | 3 | M36 | Automatic software tests and metrics and manual inspection |

| **Work package number** | **4** | | **Start - End:** | | M1 – M36 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Work package title** | **USER CENTRIC MEASUREMENT** | | | | | | | | | |
| **Activity type** | RTD | | | | | | | | | |
| **Participant number** | 1 | 2 | 3 | **4** | 5 | 6 | 7 | 8 | 9 | 10 |
| **Participant short name** | ERCIM | MAAYA | UPC | **DIALOGIC** | CNRS | KYOS | FHG | CWI | VOCAPIA | NIELSEN |
| **Person-months per participant** | | 8 | | **23** | | 4 | 2 | | | 2 |

**Objectives**

The goal of this work package is to measure actual language behaviour on the internet at the most detailed level: the individual user.

***Tangible outcomes and measures of progress and success***

For each client, information on which language is being used in which type of web-based application at a particular moment in time (day time, evening, working days, weekends) and a particular frequency. Both language use on desktop PCs and mobile devices is being monitored, the first via a browser extension (including a survey module that is used to validate the data that is automatically collected via the core client), the second via a local proxy server. The data is unique in its kind, and enables in-depth analysis of language-in-actual-use across specific background variables (gender, age, nationality etcetera), and across specific types of use settings (type of online activity, time).

A throughput measure of success of the two types of clients is the uptake (download and installation) by users, that is, the growth of the installed base of the client in various countries. The quality of the client in terms of user experiences can be judged by the churn in the panel (high churn suggests flaws in the user experience). An output measure of success is the quality and especially the uniqueness of the data when compared to the other data collection strands (WP5 and WP6). An outcome measure of success is the subsequent use of the data generated by WP4 in academic research, public policy (MAAYA; Tasks7.4/5) and commercial applications (Nielsen, Tasks7.2/3).

**Description of work**

**Task 4.1** Input specification: indicators, languages, panel, survey **(M1-M3, 8 PM)**
*Task Leader: MAAYA*
*Participants: DIALOGIC, NIELSEN*
This task consists of two parts: developing a protocol and composing a panel. Development of the protocol includes the definition and operationalization of indicators (dependent variables: language use; independent variables: relevant background variables). Three (linked) versions of the protocol will be made: one for the browser-based (desktop) version of the automated data collection, one for the server-based (mobile) version and one for the online survey.

The number and type of background variables largely determine the design of the panels. In each country and/or language panel size should be sufficiently big to enable cross-sections that generate useful results. For the browser-based measurement we aim for a minimal panel size of 1,000, in at least ten countries. For the server-based measurement we propose to target four European countries that are (de facto) bi (or multi) lingual: Ireland (Gaelic, English), Malta (Maltese, English), Luxembourg (French, German, Luxembourgish) and the Netherlands (Dutch, English).

**Task 4.2** Supported configurations specification **(M1-M3, 5 PM)**
*Task Leader: DIALOGIC*
*Participants: MAAYA, FRAUNHOFER*
The number and type of devices people use to connect to the Internet is steadily increasing lately. Most of these devices run on specific versions of operating systems. Although we can re-use some modules in practice this means we will have to develop tailor-made versions of the client for each of the versions that is involved in the project. Hence the scoping of the project is critical to the feasibility of tasks T4.5 and T4.7. This task is to specify the browser types (T4.5) and type of mobile devices (T4.7) that will be supported. Although the proxy server based measurement (T4.7) is itself platform independent the automatic installation module (to set up the default settings on the mobile device) is still platform-dependent.

At this moment (end of year 2012) we propose to consider the following configurations for the feasibility tests:

- Browsers: IE (7.0 to 10.0), Chrome/Safari (4.0-6.0), Firefox (4.0 and up)
- Mobile: iOS (4 to 6), Android (3.*,4.*), Blackberry (5, 6, 7.*), Windows Phone (7.*, 8.*)

**Task 4.3** Panel management **(M20-M31, 14 PM)**
*Task Leader: DIALOGIC*
*Participants: MAAYA, NIELSEN*
This task covers the recruitment of panellists, the distribution of the monitoring software, and the subsequent transmission of survey questions (see T4.6). User panels, of users that accept the installation of the software, will be managed here, via email and an automated multilingual help desk, or bug management system (tickets). A CRM-based system will be used to minimize administrative burden for the panellists. The system is also needed to target the recruitment of new panellists (in case of churn). We will test and use various approaches in parallel to recruit panellists, including institutional accreditation (UNESCO, possibly ITU through ISPs), raffling a price to random drawn panellists, micro-payments and/or hiring existing panels through a web marketing company.
The panel will run for one year and be dismissed afterwards.

**Task 4.4** Security and privacy audits **(M7-M31, 14 PM)**
*Task Leader: KYOS*
*Participants:*
This task will audit and address any privacy and security risks that are involved in the data exchange to and from the client of the User-centric Measurement. Four types of risks may be identified: the identification of the user based on the data that is being sent from the client; the interception of data during the exchange; the delivery of unwanted data to the client (including malevolent scripts that can be used to take over the client) and the exploitation of the security hole in the server, allowing a hacker to compromise the system. The risk of the program being affected by specific (monitoring) viruses will also be analysed and tested.
Depending on the criticality of found vulnerabilities the auditor will propose technical or organisational solutions in order to reduce the risk and work in close collaboration with User-centric Measurement participants. Once the solutions are implemented, a second set of tests and validations will be performed in order to measure the risk mitigation.

**Task 4.5** Development of the desk top based user-centric measurement client (core client) **(M2-M18, 15 PM)**
*Task Leader: DIALOGIC*
*Participants:*
Programming the monitoring client is the heart of this work package. The client is an extension to the browser that monitors the language that is being used in the web traffic. Giving the highly sensitive

nature of the data that is collected by the client, maximum care will be given to the privacy of the respondent and the security of the data. All data will be anonymised locally, at the client side, before it is being sent over an encrypted connection. Only data at a meta-level of the original texts will be transferred to and stored at the central server. Consequently, the language identification module will run at the client side, not at the server side, and should therefore be extremely light and fast to minimize use of resources from the devices on which the client is being installed.

Identification of the specific user of the desk top computer is done by manual self-identification. After installation of the client some questions will be asked via the survey module (see T4.6) to create a personal profile for each household member.

**Task 4.6** Development of the desk top based user-centric measurement client (survey module) **(M7-M19, 15 PM)**
*Task Leader: DIALOGIC*
*Participants: MAAYA*
The core client (T4.5) also has a built-in survey module. Survey questions appear in pop-ups in the client and can be triggered automatically, by specific events, or manually. We will develop a develop a web-based interface that enables user-friendly access to the (large amounts of near-real time) data. The interface is used to link the input and output from the core and survey module on the level of individual (anonymous) respondents or subsets of respondents using hash tables. Thus we can send survey questions to a specific respondent without ever knowing the identity of the respondent. Use of the survey module is strictly limited to the purposes of this research project.

**Task 4.7** Development of the user-centric measurement for mobile devices **(M2-M19, 5 PM)**
*Task Leader: DIALOGIC*
*Participants: FRAUNHOFER*
The technical challenge in the measurement of language use on mobile devices is the myriad of existing types of devices. Since each type requires a tailor-made client, it would take a lot of efforts to cover a sizeable number of devices. We have therefore opted for an entirely different solution, namely the rerouting of outgoing mobile traffic over a (local) proxy server. This solution is platform-independent (at least, the measurement part, not the installation part, see T4.2). This means that we only have to develop one generic proxy server.

To make sure that the user experience is not negatively affected (e.g., slower response times) we will set-up local servers in each of the four countries involved (see T4.1). Similar to T4.5, all data will be anonymised locally, at the client side, before it is being sent over an encrypted connection. For various protocols (e.g., HTTP, SMTP) the language-relevant part of the message will be temporary stored in the buffer and analysed. The challenge is to find an optimal balance in number and types of protocols supported, proportion of the message buffered, and processor time required for the identification of the language. Therefore, the design, set-up and maintenance of the underlying database infrastructure to store and handle the data is inherent part of this task. This infrastructure fits seamlessly into the overall IT architecture from the project as defined in T3.3 and administered in T3.4

**Deliverables** (brief description) and month of delivery

| Number | Description | WP | Nature | Diss level | Month Due |
|---|---|---|---|---|---|
| D4.1 | Report on scope of the project in terms of number and types of operating systems supported | 4 | R | PU | M4 |
| D4.2 | First stable version of protocol | 4 | R | PU | M4 |
| D4.3 | Mailing list of panel (initial set of respondents) | 4 | R | RE | M20 |
| D4.4 | Audit report on privacy and security aspects | 4 | R | PU | M19 |

| | | | | | |
|---|---|---|---|---|---|
| D4.5 | User-centric measurement clients (both desk top and mobile based), including survey module and back end interface | 4 | P | RE | M36 |
| D4.6 | Final report based on the analysis of the user-centric measurements | 4 | R | PU | M36 |

| **Milestones** | | | | |
|---|---|---|---|---|
| Number | Short description | WP | Month | Means of verification |
| M4.1 | Baseline protocol ready | 4 | M5 | Protocol has been thoroughly tested internally and tested successfully on an external panel |
| M4.2 | Baseline panel ready | 4 | M12 | A panel has been composed that meets the critical statistical size in each country involved |
| M4.3 | Beta version 1.0 of client ready | 4 | M18 | First version of the client available, technically tested (bug free) |
| M4.4 | Beta version 1.1 of client ready | 4 | M19 | Second version of the client available, with survey module integrated, technically tested (bug free) |
| M4.5 | Panel active for 6 months (halfway full duration) | 4 | M25 | Client installed at devices from 10,000 users in 10 countries, with a consistent 90% of active users |
| M4.6 | Debriefing of panel (end of duration, after 12 months) | 4 | M31 | All 10,000 users informed about the completion of the measurement. |

| **Work package number** | **5** | | **Start - End:** | | M1 – M36 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Work package title** | **ANALYSIS OF SPOKEN LANGUAGE IN MULTIMEDIA DATA** | | | | | | | | | |
| **Activity type** | RTD | | | | | | | | | |
| **Participant number** | 1 | 2 | 3 | 4 | **5** | 6 | 7 | 8 | 9 | 10 |
| **Participant short name** | ERCIM | MAAYA | UPC | DIALOGIC | **CNRS** | KYOS | FHG | CWI | VOCAPIA | NIELSEN |
| **Person-months per participant** | | | | | **14.5** | | 1 | | 16 | 1 |

**Objectives**

Much of the information on the web is not in a textual format, and therefore will escape detection, classification and categorization via text-based methods. This work package will be concerned with sampling a wide range of websites hosting audio and video documents, and developing methods to identify the languages spoken in them. A second activity will be to explore techniques to identify content categories and other information taking into account specificities of audio data.

***Tangible outcomes and measures of progress and success***

Prototype language identification systems will be developed and made accessible to the partners via a web-based service. Performance will be assessed on representative test data selected by the partners. Content categories will be annotated in the automatic transcripts for 6 languages (French, Dutch, English, Spanish, German and Italian) for which state-of-the-art speech-to-text STT systems are available (task 5.4). The automatic transcripts, provided via the web-based transcription service will also be used as text input to allow the developed opinion and sentiment analysis methods of WP6 to be applied to audio and audiovisual documents for the languages specified above.

**Description of work**

**Task 5.1** : Specification of the representative corpus obtained via large scale sampling **(M1-M6, 4 PM)**
*Task Leader: CNRS*
*Participants: FRAUNHOFER, VOCAPIA, NIELSEN*

In collaboration with the work in work packages 3 and 6, a large sample of heterogeneous audio and audiovisual corpora will be identified.

**Task 5.2** Annotation of the representative corpus with language tags **(M1-M15, 4.5 PM)**
*Task Leader: CNRS*
*Participants: VOCAPIA*
This task is concerned with the annotation of a corpus with language tags. This corpus is required for both model training and for evaluation purposes. Innovative methods will be explored to obtain the labelled data by incorporating speech technologies in the annotation process. For example, since a single audio document may contain segments in different languages, the document will first be automatically partitioned into clusters of segments nominally corresponding to a speaker in a given language, and individual clusters will be presented to humans for annotation and/or validation.

**Task 5.3** Language identification from audio **(M2-M34, 14 PM)**
*Task Leader: VOCAPIA*
*Participants: CNRS*
This task is concerned with developing models for language identification (LID) and speech-to-text

(STT). Several novel techniques will be explored to improve the quality of target language phonotactic models such as optimization of decoding parameters and intelligent selection of phone contexts; and improved decoding making use of multiple hypotheses and automatic learning techniques, and to deal with audio documents containing multiple languages.

This task will result in an LID system that will identify all UE official languages (Bulgarian, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek, Hungarian, Irish, Italian, Latvian, Lithuanian, Maltese, Polish, Portuguese, Romanian, Slovak, Slovene, Spanish and Swedish) as well as important secondary and immigrant languages (Arabic, Catalan, Mandarin, Turkish). STT systems will be developed for the Turkish and Catalan languages.

**Task 5.4** Model Development for Automatic Content Category Identification **(M6-M35, 10 PM)**
*Task Leader: CNRS*
*Participants: VOCAPIA*
This task is concerned with developing models for automatic content (topic) identification in audio data based on automatic speech-to-text transcription for covered languages. Existing speech-to-text systems for French, Dutch, English, Spanish, German and Italian will be used to transform multimedia data into texts, providing additional information such as confidences scores, and alternative hypotheses. These models will take into account specificities of automatic processing of spoken language (transcription errors, hesitations, repetitions, reformulations, etc). The topics will be defined in collaboration with the other partners and in coordination with WP6 and the use cases.

**Deliverables** (brief description) and month of delivery

| Number | Description | WP | Nature | Diss level | Month Due |
|---|---|---|---|---|---|
| D5.1 | Corpus Specification | 5 | R | PU | M3 |
| D5.2 | Report on Annotated Corpus | 5 | R | PU | M12 |
| D5.3 | Report on Language Identification in audio | 5 | R | PU | M25 |
| D5.4 | Report on Topic Detection in audio | 5 | R | PU | M35 |

**Milestones**

| Number | Short description | WP | Month | Means of verification |
|---|---|---|---|---|
| M5.1 | Baseline web-based LID service available | 5 | M6 | Automatic software tests and metrics and manual inspection |
| M5.2 | Web-based STT service available (6 languages) | 5 | M6 | Automatic software tests and metrics and manual inspection |
| M5.3 | Baseline web-based topic detection service available (8 languages) | 5 | M18 | Automatic software tests and metrics and manual inspection |
| M5.4 | Final web-based LID service available | 5 | M24 | Automatic software tests and metrics and manual inspection |
| M5.5 | Final web-based topic detection service available (8 languages) | 5 | M30 | Automatic software tests and metrics and manual inspection |

| **Work package number** | **6** | | **Start - End:** | | M1 – M36 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Work package title** | **CONTENT ANALYSIS ON MULTIMEDIA DATA** | | | | | | | | | |
| **Activity type** | RTD | | | | | | | | | |
| **Participant number** | 1 | 2 | **3** | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| **Participant short name** | ERCIM | MAAYA | **UPC** | DIALOGIC | CNRS | KYOS | FHG | CWI | VOCAPIA | NIELSEN |
| **Person-months per participant** | | | **20** | | 7.5 | | 39 | 18 | 10.5 | 6 |

**Objectives**

The work package will process textual user input from PCs and smart phones (WP4), transcribed speech from the internet (WP5), and web pages collected from the internet (Task 6.1). It will focus on the following objectives:

- Train models for the extraction of semantic information (language, document categories, disambiguated keywords, opinions, relational information) for multiple languages and text genres. Establish a real-time workflow for the collection, semantic processing and evaluation of internet documents.

Interoperability is based on common standards and exchange formats.

***Tangible outcomes and measures of progress and success***

The tangible outcomes include:

- Language identification systems able to distinguish a wide range of languages.
- Machine Translation systems able to port documents from languages with training resources for classification or opinion mining systems to languages with lack of resources, in order to build training material for those languages.
- Document classifiers for different languages ensuring classes of identical semantic over all languages.
- A keyword identifier able to find and disambiguate keywords provided by the use cases. The meaning of keywords is identical across languages.
- Beyond state-of-the-art opinion mining systems for aspect-oriented opinions able to work on a variety of languages and ensuring the same semantics across languages.
- An integrated near real-time workflow extracting semantic annotations from different modalities which ensures the same meaning of annotations for all languages.

The effectiveness of modules will be measured by software metrics and with tests on independent test sets with known result.

**Description of work**

**Task 6.1** Intelligent data gathering and language identification **(M1-M24, 10 PM)**

*Task Leader: FRAUNHOFER*

*Participants: CNRS, VOCAPIA*

This task will take care of crawling parts of the web and get a representative sample of documents the content of which will be analysed in the other tasks of the WP. This task also addresses the development of language identifiers on text, covering a wide range of languages. In addition duplicate web pages are detected .The range of websites to be crawled will be defined by the use cases in WP 7.

**Task 6.2** Prepare multilingual training data **(M1-M28, 28 PM)**

*Task Leader: UPC*
*Participants: CNRS, VOCAPIA, NIELSEN*
Documents in languages with available training data for classification and opinion mining tasks will be translated to languages with less availability of such data to develop classifiers. This task will develop MT systems for the required language pairs, gathering available training corpora for machine translation, and using domain adaptation techniques to extend their coverage. In the specific use cases new training data for text categorization and opinion mining has to be annotated using an annotation front-end.
Additionally, pre-processing functionalities will be offered for a reduced set of languages, including at least English and Spanish, in order to enrich the text with linguistic information (PoS tagging, parsing, etc.) that will enable an extended feature representation for later tasks of classification and opinion mining.

**Task 6.3** Disambiguation and classification **(M6-M30, 19 PM)**
*Task Leader: FRAUNHOFER*
*Participants: CNRS, VOCAPIA*
This task will address the development of document categorization systems for at least 9 languages. It will handle multilingualism by using translated training data or by porting target documents to a language where trained classifiers are available. Documents will be classified by genre (e.g. news, advertisement, blog, etc.) and content (e.g. politics, sports, shopping, etc.) in a classification hierarchy provided by the use cases. Depending on the type of text (web, blogs, twitter, transcribed speech) different classifiers may be required. In a similar way multilingual Named Entity Recognition will be provided.
The task will implement a multilingual keyword search to identify keywords provided by the use case which were translated to the different languages. In addition a context-based disambiguation will take place to distinguish different meanings of the keywords.
All methods will be evaluated on independent test sets to assess their performance for the different languages.

**Task 6.4** Opinion mining **(M6-M30, 26 PM)**
*Task Leader: FRAUNHOFER*
*Participants: CNRS, VOCAPIA*
This task will tackle opinion mining for at least 6 languages. On one hand, the global sentiments expressed in documents or sentences will be analyzed. On the other hand, the opinion with respect to specific keywords or with respect to aspects of these keywords is extracted. These keywords are provided by the use cases.
Task 6.2 will provide translated training data and resources for training models in different languages. In the task different supervised and unsupervised approaches will be implemented and combined if this proves to be efficient. It will be investigated whether opinions with more that two values (positive vs. negative) can be extracted.
All methods will be evaluated on independent test sets to assess their performance for the different languages.

**Task 6.5** Evaluation and aggregation of results **(M12-M36, 18 PM)**
*Task Leader: CWI*
*Participants:*
Tasks 6.3 and 6.4 provide results on a detailed level. These results are stored in an appropriate repository together with information on the document like URL, genre, date and time and regional information and have to be prepared for evaluation. If required by the use case the annotated documents will also be stored for reference and further analysis. In addition the structural relation of documents (links, followers, etc.) is evaluated if required by the use cases. In cooperation with the use cases appropriate aggregation, interactive evaluation, and presentation strategies are defined and

implemented. In addition, flexible data mining and significance analysis tools will be provided for the extraction results.

| **Deliverables** (brief description) and month of delivery | | | | | |
|---|---|---|---|---|---|
| Number | Description | WP | Nature | Diss level | Month Due |
| D6.1 | Specification of APIs and integration design | 6 | R | PU | 6 |
| D6.2 | Report on Crawler prototype | 6 | R | PU | 12 |
| D6.3 | Report on Machine Translation prototype | 6 | R | PU | 12 |
| D6.4 | Report on Linguistic enricher prototype | 6 | R | PU | 18 |
| D6.5 | Report on Disambiguation and Classification prototype | 6 | R | PU | 12 |
| D6.6 | Report on Opinion Mining Prototype | 6 | R | PU | 18 |
| D6.7 | Report on Aggregation and evaluation prototype | 6 | R | PU | 18 |
| D6.8 | Content Analysis prototype | 6 | P | PU | 30 |
| D6.9 | Report on the Workflow for the Extension of the Analysis by another Language | 6 | R | PU | 36 |

| **Milestones** | | | | |
|---|---|---|---|---|
| Number | Short description | WP | Month | Means of verification |
| M6.1 | Baseline web-based crawler service available | 6 | M12 | Tests by independent test sets with known result |
| M6.2 | Baseline web-based MT service available | 6 | M12 | Tests by independent test sets with known result |
| M6.3 | Baseline web-based linguistic enricher service available | 6 | M18 | Tests by independent test sets with known result |
| M6.4 | Baseline web-based classification service available | 6 | M12 | Tests by independent test sets with known result |
| M6.5 | Baseline web-based opinion mining service available | 6 | M18 | Tests by independent test sets with known result |
| M6.6 | Baseline web-based content analysis tools available | 6 | M18 | Tests by independent test sets with known result |
| M6.7 | Final web-based crawler service available | 6 | M24 | Tests by independent test sets with known result |
| M6.8 | Final web-based MT service available | 6 | M24 | Tests by independent test sets with known result |
| M6.9 | Final web-based linguistic enricher service available | 6 | M28 | Tests by independent test sets with known result |
| M6.10 | Final web-based classification service available | 6 | M30 | Tests by independent test sets with known result |
| M6.11 | Final web-based opinion mining service available | 6 | M30 | Tests by independent test sets with known result |

| **Work package number** | **7** | | **Start - End:** | | M1 – M36 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Work package title** | **USE CASES** | | | | | | | | | |
| **Activity type** | RTD | | | | | | | | | |
| **Participant number** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | **10** |
| **Participant short name** | ERCIM | MAAYA | UPC | DIALOGIC | CNRS | KYOS | FHG | CWI | VOCAPIA | **NIELSEN** |
| **Person-months per participant** | | 30 | | 3 | 2 | | 10.5 | 1 | 2 | **18** |

**Objectives**

WP7 plays a key role in the project in applying the methods and tools created by SEMACORE research work packages into a large and diverse range of real life situations (from companies, public agencies and international organizations contexts) in order to prove its concrete value. Together with this proof of concept the objective is to assess the products and methods and forecast future impacts.

***Tangible outcomes and measures of progress and success***

The tangible outcomes includes:

- Prototypes for three use cases covering individual and diverse situations of real life use of the SEMACORE tools and methods which will be evaluated
- Evaluation from proven data of the production of the project in terms of indicators
- An impact forecast model

**Description of work**

**Task 7.1** Linguistic diversity use case **(M1-M36, 13 PM)**

*Task Leader: MAAYA*

*Participants:*

The SEMACORE research results and methods will be used in concrete case studies to evaluate characteristics of the contents of global digital libraries, international organizations websites, the whole web space of four European countries and a selection of European blogs reflecting citizen opinions.

- Google Books, Europeana, ScholarVox and World Digital Library will be analyzed for duplicated records and for the identification of contents by language.
- EC and UN web sites will be analyzed to obtain a picture of their respective language usage proportions.
- Full characterization of the national web of Ireland, Luxembourg, Malta and Netherlands will be performed. The ccTLD of each country will be crawled systematically and a set of data will be collected following traditional parameters (language per page, type of domain per web, size of web site, links in/out…). In addition, content categorization, as well as automatic quality criteria (e.g. derived by bootstrap) will be established at the page level, allowing an innovative characterization. Cooperation with National Information Centers (NIC) will be organized to facilitate the management and whereas possible obtain the table of IP numbers of the country allocated to generic domains so to extend the characterization. For countries having too large a web domain for systematic crawling, some methods will be employed (such as streaming sampling) in order to allow the characterization from a subset in the order of magnitude of less than 10% of the total. Finally a guideline document will be established to describe the steps towards country web characterization and the challenges and solutions, as well as the lessons

learned during the use case.

- A sample of European blogs in English, French and German will be selected to analyze citizen sentiments as regard to online use of languages and learning languages. The question to be analyzed will be drawn and extended from Eurobarometer surveys[44] focusing languages.

**Task 7.2** Content and opinion mining for an international public agency **(M1-M36, 13 PM)**
*Task Leader: NIELSEN*
*Participants: FRAUNHOFFER, CNRS, VOCAPIA*
For the European Commission DG SANCO: SEMACORE will collect web data for a special topic (e.g. Health or Hydration) at different time points in different countries. Across different languages occurrences of a number of predefined key phrases (e.g. drug names, health issue names, or EU initiatives). The application will analyse the distribution of keywords over different media, languages, and content categories and describe their temporal evolution. In addition relevant aspects of keywords (e.g. risk, effectiveness, price, or physiological compatibility) will be extracted. The analysis will also include audio and video sources using Speech-To-Text or print media to compare statements collected from those sources. Finally opinions (positive, neutral, negative) on keywords and attributes/aspects will be extracted across languages. Keywords and aspects will be grouped into a hierarchical ontology.
A flexible evaluation GUI will be provided which allows the joint analysis of keywords and opinions and the corresponding categories, languages, etc. Using past documents stored in the repository the dynamic development of keyword mentions and opinions can be analysed and "first occurrences" of words or phrases may be determined. There will be a close cooperation with the customer to perform a demand driven search and evaluation process. The use case will not only serve as a means to support policy formulation and the monitoring of ongoing efforts but may also be used to inform the general public timely and thoroughly on the attitude of citizens with an effort to improve e-Participation and the inclusion of citizens in the policy process.

**Task 7.3:** Content and Opinion mining for an international company **(M1-M36, 16 PM)**
*Task Leader: NIELSEN*
*Participants: FRAUNHOFFER*
This use case will be conducted for an international company (e.g. Fast Moving Consumer Good company). It will be conducted for a product or service including a full analyse of the keywords, categorisation of them, sentiment and distribution of the data and results within association maps, category information, as well as impact analysis including recommendations. Nielsen will provide extensive training data for a hierarchy of 150 use case specific categories. Documents are collected by focussed crawling and are analysed to detect occurrences of a number of predefined key phrases (e.g. disease names, issue names, or names of competitors and products). The texts will be classified into a hierarchical ontology. Relevant named entities (e.g. product names, product features) and aspects/attributes of keywords will be extracted. The application will analyse the distribution of keywords and their aspects over different media, languages, genres and content categories, and the temporal evolution. Relevant data from external sources such as Speech To Text or print media will be integrated. Furthermore the opinions (positive, neutral or negative) expressed on specific keywords and on relevant aspects will be extracted. Relevant parts of the collected documents will be stored in the content repository for the analysis of dynamic developments and retroactive searching for "first occurrences" of words or phrases. A use case specific GUI will be provided to allow interactive presentations of results tailored to the needs of the customer.

**Task 7.4:** Evaluation of language indicators **(M1-M36, 20 PM)**
*Task Leader: MAAYA*
*Participants: FRAUNHOFFER, CWI, DIALOGIC*
The activity consists in adapting and applying the words sampling method of FUNREDES with the new

[44] http://ec.europa.eu/public_opinion/archives/ebs/ebs_386_en.pdf

elaborated strategies for crawling as produced by WP3 in order to establish world wide figures of languages presence for a limited set of languages. The product will serve as input to assess new methods produced by the project and offer a reliable reference with historical series for crosschecking figures.

**Task 7.5** Use cases evaluation and future impact forecast **(M1-M36, 5 PM)**
*Task Leader: MAAYA*
*Participants:*
The activities consist in the creation of a methodological framework for the Uses Cases to both, evaluate the impacts obtained by the research which have been conducted as well as the results which have been obtained and forecast future impacts of the project. In that sense, T7.5 will act both as front-end and back-end to the other tasks of WP7 and play a transversal role to help emerge coherence and synergy from the different parts of the projects viewed from the users and uses cases point of view. The reports produced by T7.5, together with those of T8.2 for a sustainability roadmap, will provide valuable insights for future research and applications driven from the SEMACORE project as well as present a comprehensive and coherent balance of the outcomes of the project as they impact the various fields involved (analytics in big data, voice technologies, linguistic diversity, information society indicators, content industry, public policies for digital divide).

**Deliverables** (brief description) and month of delivery

| Number | Description | WP | Nature | Diss level | Month Due |
|---|---|---|---|---|---|
| D7.1 | Initial progress report on SEMACORE Use Cases covering all sources described within WP7 | 7 | R | PU | 12 |
| D7.2 | Intermediary progress report on SEMACORE Use Cases | 7 | R | PU | 24 |
| D7.3 | Final progress report on SEMACORE Use Cases | 7 | R | PU | 36 |
| D7.4 | Report on Funredes methods measurements and cross-checking and validation of project produced data | 7 | R | PU | 36 |
| D7.5 | Strategy paper on forecast impact | 7 | O | PU | 36 |
| D7.6 | Report on Use Cases evaluation | 7 | R | PU | 36 |

**Milestones**

| Number | Short description | WP | Month | Means of verification |
|---|---|---|---|---|
| M7.1 | Use Cases planning | 7 | M12 | Report evaluated by EAB |
| M7.2 | Software programming and test of T7.4 | 7 | M18 | Test success |
| M7.3 | Uses cases launch | 7 | M18 | Partner meeting |
| M7.4 | Uses cases completion | 7 | M30 | Reports |
| M7.5 | Uses case evaluation and impact forecast | 7 | M36 | Reports |

| **Work package number** | **8** | | **Start - End:** | | M1 – M36 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Work package title** | **DISSEMINATION AND EXPLOITATION** | | | | | | | | | |
| **Activity type** | OTH | | | | | | | | | |
| **Participant number** | 1 | **2** | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| **Participant short name** | ERCIM | **MAAYA** | UPC | DIALOGIC | CNRS | KYOS | FHG | CWI | VOCAPIA | NIELSEN |
| **Person-months per participant** | 4 | **18.5** | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 3 |

**Objectives**

This work-package is conceived as the main guarantee of the return on the research investment into visible and exploitable products with managed impacts and the corresponding dissemination and exploitation of both the research results and the impacts, among all relevant stakeholder categories. The dissemination will be realized both directly in a comprehensive web site reflecting the progress and outcomes of the research and indirectly through appropriate publications targeting the research and the business communities as well as public at large. Particular attention is given to the economic exploitation, and the provision of a standardized workflow for including new languages in the joint analysis. Two workshops will be organized to consolidate the strategic directions of the project with relevant stakeholders, respectively on the business field and on the policy field. Finally there will be a roadmap to warrant the sustainability of the language indicators production.

***Tangible outcomes and measures of progress and success***

The tangible outcome of this work package encompasses:

- a web site which shall become a reference on the processed subjects and which measure of progress will be the incoming traffic and the number of links to.
- a set of publications covering the results of the research realized, the impact of the project in terms of e-business; the success of the publication will be measured by the number of citations;
- an exploitation strategy based on a market analysis which will be documented and assessed by the External Advisory Board.
- a strategy for the sustainability of the indicators production which will be assessed by the External Advisory Board.
- a Roadmap for Internet linguistic diversity measurement which will be assessed by the External Advisory Board.

**Description of work**

**Task 8.1** Web site and production of promotional material **(M1-M36, 4 PM)**

*Task Leader: ERCIM*

*Participants: (all)*

Overall design, creation, hosting and maintenance of the project web site in various releases and production of a set of promotional material (brochure, flyers, posters, presentations) which will be used by all the partners attending events related to the topics of the project. There will be at least 5 press releases on the application scenarios.

**Task 8.2** Institutional exploitation and sustainability roadmap **(M1-M36, 18.5 PM)**

*Task Leader: MAAYA*

*Participants:*
This task addresses the issue of the sustainability of the production of indicators for institutional partners. It will document the steps necessary to add a new language to the joint analysis workflow. The agreements which will be required with some of the partners for future production of indicators in a non business context will be documented. The tasks provides policy and research recommendations, aiming at making sure that the research results of the project are taken up by the appropriate research communities and have an impact on policy in the field beyond the project lifetime. A foresight exercise will be developed to consider future directions for R&D to measure linguistic diversity in the digital world. The topics to be investigated will include: relationship between research, policy and practice; identifying common challenges and key issues facing programmes; ideas for future R&D co-operation; new opportunities and gaps in R&D as well as the possibility to set up a future stakeholder network and make recommendations for future collaboration activities with similar and complementary projects, clusters and programs. A workshop will be organized with the relevant stakeholders to consolidate the findings of the foresight exercise. Based on the results of this foresight exercise and the outcomes of the workshop, a Research Roadmap for internet linguistic diversity measurement will be produced.

**Task 8.3** Academic and scientific dissemination **(M1-M36, 0.9 PM)**
*Task Leader: FRAUNHOFER*
*Participants: UPC , CNRS, CWI*
Providing inputs for the research section of the web site and coordination of a set of scientific publications reporting to the research community on the research outcomes of the project. The technical partners and user partners are strongly committed to dissemination of results through well-recognized scientific channels. All have excellent records of publication at major scientific venues in their respective fields, conferences (ISCA Interspeech and workshops, IEEE ICASSP, KDD, ASRU, LREC, HLT, ECML/PKDD, ACL, CLEF, ACM SIGIR, etc) and journals (Speech Communication, Computer Speech and Language, Natural Language Engineering, Machine Learning, Data Mining, Information Retrieval, etc). The interdisciplinary nature of the project will provide an important opportunity for the partners to inform experts in fields beyond their own about the impact of the work and the collaboration.

**Task 8.4** Commercial dissemination and Exploitation **(M1-M36, 4.2 PM)**
*Task Leader: NIELSEN*
*Participants: DIALOGIC, VOCAPIA, KYOS*
This task will create a credible exploitation strategy based on a market analysis. It will determine the tactical steps required for implementing the commercial exploitation strategy and will identify the major SEMACORE results and describe how these results are exploited by the consortium. Nielsen as industrial partner will coordinate the effort to exploit the results of the work in this project by using the developed platform to improve and create new solutions in the area of language and content analysis to provide insights into the elicitation of the sentiments and opinions of citizens on political issues and economic topics such as the reputation of products and enterprises. There will be presentations of the business oriented applications in appropriate venues such as Online Information, WWW Conference, Content Management, Semantic Technologies, IFRA, Social Media World Forum, etc.
This task will organize a high-quality scientific workshop. The goal of the workshop is to bring together researchers, developers, end-users, and practitioners carrying out research and development in this area and provide a forum for discussing scientific and practical issues covering all aspects of conceptual, design and use related to the SEMACORE project. We will investigate the possibility of organizing the workshops at top-tier conferences such as VLDB, ICDE, ECML/PKDD, ICDM or ICML. The aim is to widen the research and application community and to discuss and compare with related work.

**Deliverables** (brief description) and month of delivery

| Number | Description | WP | Nature | Diss level | Month Due |
|---|---|---|---|---|---|

| | | | | | |
|---|---|---|---|---|---|
| D8.1 | Initial project web site | 8 | O | PU | M3 |
| D8.2 | Dissemination and use plan V1 | 8 | O | PU | M6 |
| D8.3 | Dissemination and use plan V2 | 8 | O | PU | M18 |
| D8.4 | Strategy document for research roadmap | 8 | O | PU | M30 |
| D8.5 | Strategy document for indicator sustainability | 8 | O | PU | M30 |
| D8.6 | Final dissemination and exploitation report | 8 | O | PU | M36 |

| **Milestones** | | | | |
|---|---|---|---|---|
| Number | Short description | WP | Month | Means of verification |
| M8.1 | Consolidated version of web site | 8 | M12 | Online review |
| M8.2 | Consolidated plan for all dissemination and exploitation activities | 8 | M18 | Partner meeting |
| M8.3 | Review of plan for all dissemination and exploitation activities | 8 | M24 | Partner meeting |
| M8.4 | Workshop for strategies on business | 8 | M30 | Web site |
| M8.5 | Workshop for strategies on policies | 8 | M30 | Web site |
| M8.6 | Evaluation of all dissemination and exploitation activities | 8 | M36 | Report |

### B 1.3.9 Summary of effort

The following table provides a summary of the planned effort, for each work package by each participant in person-months. The work-package leader for each WP is identified by showing the relevant person-month figure **in black**.

| Partic. no. | Partic. short name | WP 1 | WP 2 | WP 3 | WP 4 | WP 5 | WP 6 | WP 7 | WP 8 | Total PM per Partner |
|---|---|---|---|---|---|---|---|---|---|---|
| **1** | ERCIM | **14** | | | | | | | 4 | 18.0 |
| **2** | MAAYA | | **15** | 3 | 8 | | | 30 | **19** | 74.0 |
| **3** | UPC | | | 2 | | | **20** | | 0 | 22.3 |
| **4** | DIALOGIC | | 1 | 2 | **23** | | | 3 | 0 | 29.3 |
| **5** | CNRS | | | 3 | | **15** | 8 | 2 | 0 | 27.3 |
| **6** | KYOS | | 8 | | 4 | | | | 0 | 12.3 |
| **7** | FRAUNHOFER | | | **35** | 2 | 1 | 39 | 11 | 0 | 87.3 |
| **8** | CWI | | | 2 | | | 18 | 1 | 0 | 21.3 |
| **9** | VOCAPIA | | | 2 | | 16 | 11 | 2 | 0 | 30.3 |
| **10** | NIELSEN | | | 2 | 2 | 1 | 6 | **18** | 3 | 32.0 |
| **Total PM per WP** | | 14.0 | 24.0 | 49.5 | 39.0 | 32.5 | 101.0 | 66.5 | 27.6 | 354.1 |

**Figure 13 - Summary of effort at WP and Partner levels**

## B 1.3.10 Risk analysis

### B 1.3.10.1 Technology-related risks overview

| Risk | Description | Contingency plan or mitigation strategy |
|---|---|---|
| R1.1 | Poor coverage of crawls. The SEMACORE crawling strategies will be specified by the use case partners. They may miss relevant parts of the web. | The risk is mitigated by two factors. Firstly, in SEMACORE we may also exploit strategies based on other data e.g. the Common Crawl available to the partners. Secondly, sampling mechanisms could be adapted to overcome the limitation that crawling might incur in. |
| R1.2 | Disambiguation of keywords in different languages is not reliable. | In this case the multilingual topic models used in the disambiguation have to be improved by using more training data and perhaps more detailed topics. In addition the relation between named entities may be employed to improve disambiguation. |
| R1.3 | Cross-lingual content classification is inadequate. | This may be caused by insufficient training data or inadequate translation of training data. In the first case additional training data has to be annotated. In the second case the machine translation methods have to be improved. It is also possible to translate all documents to a single language for classification. |
| R1.4 | Cross-lingual opinion mining results are inadequate. | A possible reason is that the underlying topic models do not capture the right features. In this case the topic models and deep learning models may be improved by further training data. As a fallback documents could be translated to a single language before extracting opinions. Finally simple models based on patterns and opinion lexicons may be used. |
| R1.5 | Unable to locate sufficient appropriate audio data | Explore alternative sampling schemes with WP5 |
| R1.6 | Unable to obtain a consensus on what annotations (language / dialect / content /opinion) are needed for all applications | Define a common subset of annotations and extensions for particular uses |
| R1.7 | Unable to obtain sufficient annotations with crowd sourcing | Contact language schools or other associations to locate annotators |
| R1.8 | Automatic speech recognition is insufficient for content detection. | A small amount of data will be manually corrected for a subset of the most important languages for the project |
| R1.9 | Interactive tool remains too abstract to use in daily life analyst | Researchers create task- and context- specific expressions for the analyst's needs on the basis of interviews and observation study of analyst at work. |
| R1.10 | User accepting plug-in software does not constitute a large enough panel | International Organizations involved in the project will contribute to promotion. |
| R1.11 | Origin of users accepting plug-in software is not varied enough and provokes statistical bias. | Statistical methods will be used to correct the panel biases. |

### B 1.3.10.2 Non-technology-related risks overview

| Risk | Description | Contingency plan or mitigation strategy |
|---|---|---|
| R2.1 | Partnership risks: a partner leaves the consortium.<br><br>Commercialization at risk: industrial partner leaves the market. | In that case, we will promptly identify and recruit a new partner. A specific plan will be defined to bring the new partner up to speed in the shortest time as possible, in order to not cause further delays.<br>In that case, it is possible that a re-allocation of some tasks within the remaining consortium occurs. |
| R2.2 | Defaulting partner: a partner does not deliver his contribution to the work plan, creating severe gaps and delays. | If this occurs, it is likely to be quickly detected by the SCO (task T1.2, Resource Management that includes the assessment of individual partners). If no corrective action is undertaken by the defaulting partner, the GA has the possibility to suspend payments and to exclude the partner. |
| R2.3 | Management risk: insufficient management activity.<br><br>Inadequate communication between partners. | Additional resources from ERCIM and FRAUNHOFER will be devoted to project management. Corrective actions will be defined and implemented with the control of the PEB. |
| R2.4 | Conflict within consortium :<br>There is a severe conflict between 2 or more partners of the SEMACORE consortium that prevents any constructive collaboration. | There is an escalation procedure foreseen in the Consortium Agreement for addressing these situations. Successively, the PEB and the GA may intervene.<br>If no long term solution can be found, then the consortium will take the necessary measures, similarly than in the case of defaulting partners. |

Additional risks may be elicited, such as:

- Market risks: nobody is interested by the SEMACORE platform.
- Regulation/safety risks: new laws and regulations may limit the exploitation potential of the technology.

These risks relate to situations that may occur after the project ends. They will be appropriately managed in due time.

# B 2. Implementation

## B 2.1 Management structure and procedures

### B 2.1.1 Management structure

***Project management approach***

The SEMACORE management structure is designed to drive the efficient implementation of the Project's activities as defined in the work plan and the achievement of its scientific objectives, as well as the completion of its contractual obligations vis-à-vis the European Commission, in compliance with detailed rules and procedures.

We recognize that for Projects involving geographically separated participants, a major risk for failure is the lack of coordination. Accordingly, our Project management approach consists of a comprehensive Project management plan together with clear reporting procedures to ensure efficient Project quality and cost control, as well as visibility for all Project partners and external contacts throughout the lifetime of the Project. Special attention is brought into linking all Project components and maintaining smooth communications between the partners.

A management structure that assures close control has been established and well-defined objectives have been set for all partners to ensure agreement even before the Project begins.
Project management in SEMACORE relies on three principles:

- The creation of a simple and effective management structure to allow for quick decisions.
- The establishment of simple mechanisms to efficiently resolve problems and potential conflicts.
- Responsibilities are clearly defined for self-contained subsets of work to minimize overall risks.

***Overall structure***

The following diagram presents the overall structure of the SEMACORE management and reporting structure.

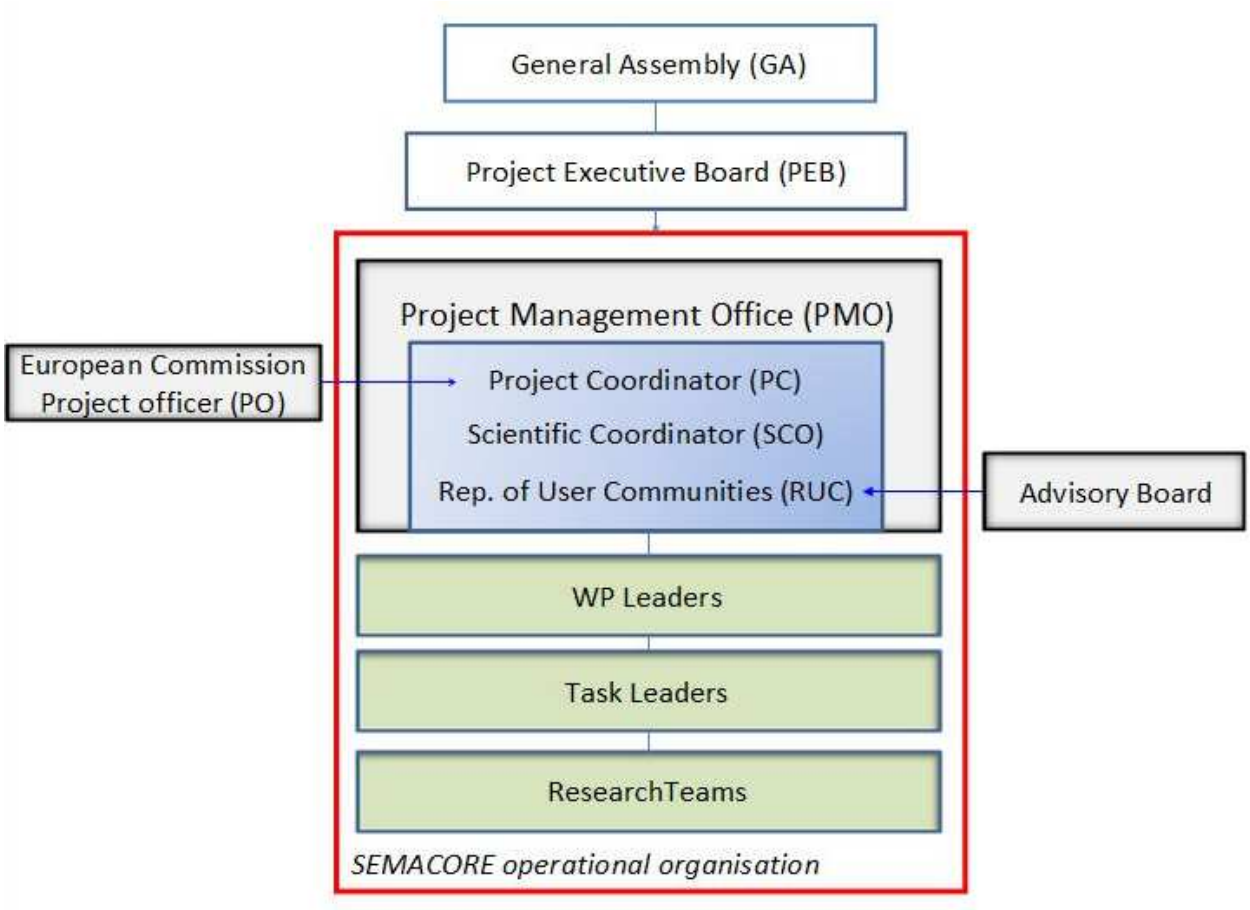


**Figure 14 - Management structure**

The boxes within the red box consist of the SEMACORE operational organisation.
The responsibility of each body or role in the organisation is defined in the following paragraphs.

***General Assembly (GA)***
The General Assembly is the ultimate decision-making body of the Consortium.
The General Assembly consists of one representative (GA Members) of each Consortium partner. Members of the GA will be designated at the start of the project.
The GA is chaired by the Project Coordinator, who initiates an ordinary meeting at least once a year and an extraordinary meeting at any time upon request of the Project Executive Board or of 1/3rd of the Members of the General Assembly.
Each GA Member present or represented in the meeting shall have one vote. Decisions shall be taken by a majority of two-thirds (2/3) of the votes. A decision is being escaladed at the GA level (in administrative, financial, scientific and technical domains) whenever the decision impacts the Project objectives or when the Project Executive Board cannot reach a consensus.

The following decisions shall be taken by the GA only:

- Evolution of the Consortium Agreement, premature termination of the Project.
- Evolution of the consortium (entry of a new partner, termination of a partner's participation).
- Evolution of the Project that impacts the grant with the commission or the content of the Project as defined in the submission documents.
- All budget-related matters, content, finances and intellectual property rights.

The GA decisions are binding to all partners in all Project-related matters. Recommendations provided by the Advisory Board, by representatives of the European Commission, and by other Project-related panels will be considered within the decision-making process.

***Project Executive Board (PEB)***

The Project Executive Board is the supervisory body in charge of the execution of the Project which reports to and is accountable to the General Assembly.

The Executive Board shall consist of the Project Coordinator, the Scientific Coordinator, and Work Package and Task Leaders as appointed by the General Assembly. The Scientific Coordinator (SCO) shall chair all meetings of the Project Executive Board. The SCO will organise ordinary PEB meetings at least quarterly and extraordinary meetings at any time upon request of any Executive Member. PEB meetings can be organised via conference calls.

Given the definitions above, the PEB will be composed of the following 11 members:

| PEB member | Organisation | WP leader of | |
|---|---|---|---|
| Philippe Rohou | ERCIM | WP1 - Project Management | Project coordinator |
| Daniel Prado<br>Daniel Pimienta | MAAYA | WP2 - Societal Issues<br>WP8 - Dissemination & Exploitation | |
| Lluis Padro | UPC | WP6 - Content analysis on multimedia data | |
| Robbin te Velde | DIALOGIC | WP4 - User-centric Measurement | |
| Lori Lamel | CNRS | WP5 - Analysis of spoken language in multimedia data | |
| Fabien Jacquier | KYOS | | |
| Gerhard Paass | FRAUNHOFER | WP3 - System Development & Integration | Scientific Coordination |
| Arjen Devries | CWI | | |
| Bianca Veru | VOCAPIA | | |
| René Lamsfuss | NIELSEN | WP7 - Use Cases | |

**Figure 15 - Project Executive Board**

The PEB shall:

- Manage and monitor the effective and efficient implementation of the Project according to the decisions of the GA and prepare the meetings of the GA when new decisions are required.
- Initiate, coordinate and have organised the Work Packages according to the Project work plan.
- Support the SCO in the scientific and technical validation of Project results and external deliverables.
- Support the PC in preparing meetings with the European Commission and in preparing related data and deliverables.
- Support the SCO in preparing meetings with the Advisory Board.
- Make the decisions related to minor changes in the Work Packages and in the technical roadmap of SEMACORE (including restructuring Work Packages if and when required).

The PEB shall seek a consensus among the partners.

***Project Management Office (PMO)***

The Project Management Office consists of a team grouping the Project Coordinator (PC), the Scientific Coordinator (SCO) and the representative of the user community (RUC). Philippe Rohou (ERCIM) will act

as the Project Coordinator; Gerhard Paass (FRAUNHOFER) will act as the Scientific Coordinator; Daniel Pimienta (MAAYA) will represent the interests of the end-users. Together they have an extensive qualification, expertise and experience relevant to Project management.

The Project Coordinator shall be the intermediary between the Partners and the European Commission and shall perform all tasks assigned to it as described in the EC Grant Agreement and in the Consortium Agreement. The Project Coordinator is in charge of the administrative and financial management of the Project. In particular, the PC shall be responsible for:

- monitoring compliance by the Partners with their contractual obligations;
- setting-up a collaborative work environment accessible to all Partners;
- keeping the address list of Members and other contact persons updated and available;
- collecting, organising a quality review process and submitting reports and other deliverables (including financial statements and related certifications) to the European Commission;
- transmitting documents and information connected with the Project to and between work package Leaders, as appropriate, and any other Partner concerned;
- chairing the SEMACORE General Assembly;
- organising the Project review meetings in coordination with the EC Project Officer;
- administering the Community financial contribution and executing the payments to Partners;
- providing, upon request, the Partners with official copies or originals of documents which are in the sole possession of the Coordinator when such copies or originals are necessary for the Parties.

The Scientific Coordinator (SCO) is in charge of the planning, management and monitoring of the research and technological development activities, including the coordination of scientific and technical work between work packages. In particular, the SCO shall be responsible for:

- monitoring the Project progress on a day-to-day basis for continuous rating of the achievements, objectives, tasks, work packages, and the entire Project;
- assessing the scientific contribution of each individual Project Partner;
- ensuring a smooth and efficient collaboration of all Partners;
- chairing the SEMACORE Project Executive Board;
- keeping close contact with the chair person of the Advisory Board;
- leading the scientific dissemination activities;
- checking the delivery of documents and information regarding the SEMACORE Project within the agreed time;
- driving the process for updating the description of work according to Project, science and technology evolution.

***Work package Leaders*** are responsible for the coordination and monitoring of all the activities composing their work package and for the liaison with the Project Management Office and the other work package leaders. They will organise intra- and inter- work package meetings as required, using extensively dedicated tools ranking from mailing lists to audio-conferencing services. They will promptly report to the PMO any deviation with respect to the Project plan in order to implement fast corrective actions.

***Task Leaders*** are responsible for the coordination and monitoring of all the activities composing their task and for the liaison with the work package leader and the other task leaders within their work package. They will organise task meetings as required, using extensively dedicated tools ranking from mailing lists to audio-conferencing services. They will promptly report to the work package leader any deviation to the Project plan.

***External Advisory Board (EAB)***

Considering that the current project has the potential to profoundly impact society and/or industry it is important that large cultural players and opinion leaders are involved. The SEMACORE Advisory Board

will be composed of representatives of such companies and institutions. The following stakeholders have already accepted to sit on the EAB:

- Gregory Grefenstette, EXALEAD, Chief Science Officer, who also accepted to chair the EAB,
- Ricardo Baeza Yates, Yahoo Research Europe, Vice-President
- Mrs Vanessa Gray, ITU, Senior ICT Analyst (ICT Data & Statistics)
- Mikami Yoshiki, , Professor Nagaoka University of Technology, leader of Language Observatory Project
- Joseph Mariani ,CNRS , director of the Institute for Multilingual and Multimedia Information
- Stephane Boyera, W3F (Web Foundation), Lead Programme Manager,
- Alexandre Wolff, OIF, Responsable de l'Observatoire de la langue française

The Advisory Board will have four specific missions:

- Monitor project progress on achieving major objectives.
- Foster innovation: preparing the adoption of SEMACORE resulting technologies by the different stakeholders, by providing input regarding the required outcome parameters, the usability, application scenarios and opportunities.
- Legal and ethical guidance: providing guidance to SEMACORE researchers to ensure that activities and technologies (carried out during the project or considered for the further exploitation of project results) are respecting European legislation and ethical principles.
- Upon request, evaluate and assess strategic reports produced by the SEMACORE work packages.

The Advisory Board will meet twice (1-day face-to-face meeting): once at the beginning of year 2, once at the beginning of year 3. More frequent interaction (by email and/or audio conference) may be set up, depending on project needs and on Advisory Board members' interest.
A "Memorandum of Understanding" will be put in place, defining the rights, duties and expectations between the Advisory Board members and the SEMACORE consortium.

***Dissemination and Exploitation Manager***

Since exploitation and dissemination tasks are essential for the success, the usage and the acceptance of the projects outcomes, a Dissemination and Exploitation Manager (DEM) will be in charge of the coordination of such tasks in the project. This will be the role of the WP8 leader (Daniel Pimienta, MAAYA). Concerning dissemination he will be responsible for all the issues related to the wide diffusion of the project results in order to ensure the largest possible visibility and successful exploitation.
At the meetings of the Project Executive Board, the DEM will introduce the dissemination agenda and address all the issues related to its attributions (publication, participation to professional and public events, awareness raising activity). Concerning exploitation, he will be in charge of liaison with all other partners dealing with exploitation, negotiation with external partners, and IPR (patents, licensing, and royalties).

***Quality management***

Quality management is of the highest importance for maximising the chances to reach all Project objectives. Quality management is driven by the Project Management Office. On the administrative level, the PC will set up and organise the quality assurance process for all Project deliverables. On the scientific and technological level, the SCO will assess the quality of the contribution of all partners and of Project results.

***Intellectual Property Rights (IPR) management***

Proper IPR management aims at setting the basis for a successful exploitation of the Project results. The general rules and procedures related to IP are defined in the Project Consortium Agreement. Fabien Jacquier (KYOS) will act as the IPR manager and will be in charge of driving the IPR-related processes, possibly setting up an IPR Committee to resolve specific issues.

### B 2.1.2 Procedures and tools

The following procedures and tools will be used in order to ensure efficient management and communication throughout the whole Project organisation.

***Management notes***
Project specific processes (e.g. deliverable review, publication approval) will be documented in Management Notes that will be produced by the PMO. Management notes will be written by the PC and approved by the SCO before being posted on the project wiki.

***Collaborative Tools***
The PMO will set up a collaborative working environment, consisting of following items:

- a set of Mailing Lists, to ensure an efficient communication (including archiving of messages) between the different communities of the project (GA, PEB, PMO, EAB, WP Leaders, etc.).
- a Collaborative web environment (BSCW or equivalent) to be a repository for all information generated by the Project, such as contractual documents, project deliverables, minutes of meetings, dissemination material, etc.
- a wiki-based discussion forum and dynamic information system to allow partners to exchange on a regular basis on current topics of interest.

***Meetings***
It is expected that the Project Executive Board will organise a monthly audio call and will meet face-to-face every quarter (one of which to prepare and attend the annual EC review). The Advisory Board will meet face-to-face once a year, in parallel with one of the PEB meetings. It is assumed that because all project partners will be invited to attend all the face-to-face PEB meetings, any of these quarterly PEB meeting can designated as the annual GA meeting.

### B 2.1.3 Conflict resolution

Even if the preferred decision making process is aimed at building consensus between the partners, a divergence or a conflict between several parties may arise. The following topics may cause such conflict:

- Technical discussion unresolved,
- Task allocation,
- Partner not delivering,
- conflict between persons,
- Funding distribution,
- IPR,
- Etc.

The SEMACORE overall structure provides a clear way to manage conflicts, and with respect to the escalation of conflict resolution (up to GA), the Consortium Agreement describes the process for settlement of disputes. The escalation process is defined as follows:

- Search for a solution by the Project Management Office,
- If no resolution is found, then involve the Project Executive Board, which is empowered to decide on minor issues,
- If no resolution is found, then involve the General Assembly, which is empowered to decide on major issues,
- Whenever appropriate, maintain a close communication with the EC Project Officer.

In order to ensure an efficient operational management of the Project activities, it is very important that the PMO expresses a unique point of view: the PC, the SCO and the EAB chair will therefore maintain a very close communication channel and seek consensus between them. The risk of conflict between the

PC and the other two coordinators is minimal because (i) the role of each one is clearly defined and (ii) the nature of their organisation makes it unlikely to have a conflict of interest. A sane and fruitful tension may arise from the existence of both a SCO and a SMC. This healthy tension will result in a pressure on researchers to deliver the data closest to what users expect on one hand, and will produce a better understanding from the users of the difference between the field of possible and the field of wishes, on another hand. Because either coordinator has sufficient knowledge of the remit of the other (research vs. indicators), we anticipate a smooth and creative decision process concerning the requirements. If ever required, the PC will play a role of arbitration.

## B 2.2 Individual participants

### B 2.2.1 *Partner 1: ERCIM, France (Coordinator)*

The **European Research Consortium for Informatics and Mathematics** (ERCIM, www.ercim.eu) is a Consortium of two organisations, a European Economic Interest Grouping (EEIG), and a Non Profit International Association (AISBL), composed of a network of research institutes from twenty two European countries, embodying more than 12,000 researchers and engineers. ERCIM is based in Sophia Antipolis (France) with an antenna in Brussels.

ERCIM's mission is to: foster collaborative work within the European research community in Information and Communication Technologies (ICT) and Applied Mathematics; advise the European Commission and national governments; and increase co-operation with European industry. ERCIM is also the European host of the World Wide Web Consortium (W3C), whose mission is to lead the World Wide Web to its full potential by developing protocols and guidelines that ensure long-term growth for the web.

***Role in SEMACORE***
ERCIM will lead WP1 and provide the financial and administrative coordination of the project. It will also contribute to the dissemination activities, in particular via the ERCIM publication, as well as leveraging its web team for the development and maintenance of the project web site. ERCIM will also call upon its W3C staff for punctual support with specific issues related to language standardisation and normalisation.

***Expected outcome from SEMACORE***
Through SEMACORE, ERCIM is fully accomplishing its mission, supporting the European leading academies and industries in their respective search for scientific and business excellence.

***Key personnel***

**Philippe Rohou** will serve as the Project Coordinator of SEMACORE, and will personally lead all tasks under ERCIM responsibility. Philippe's European project management experience includes the administrative and financial coordination of the DELOS NoE (60 partners), of the CoreGRID NoE (42 partners), of the RACE-network RFID thematic network (25 partners), of the D4Science and iMarine I3's, of the Digital World Forum and WAI-AGE CSA's, of the AXES IP, of the Net-WMS STREP, of the ABCDE Cofund project, and a few others, as well as the dissemination work package leadership of the VPH NoE.

### B 2.2.2 *Partner 2: Réseau Mondial pour la Diversité Linguistique, CH (MAAYA)*

An initiative that came out of the second phase of the World Summit on the Information Society (WSIS) in Tunis in November 2005, the World Network for Linguistic Diversity, MAAYA (http://maaya.org), aims to value linguistic diversity as a building block of uniqueness of human communications. Gathering the

most competent bodies in the field of languages and cyberspace, MAAYA serves as a multi-stakeholder network in the area of shared knowledge, where technology offers a great potential for languages, but is also a risk to them. MAAYA is a focal point for linguistic research projects and its objectives includes the promotion of software localization, equal access of all languages to cyberspace and the observation of the implementation of language policies. Amongst the founding members of MAAYA are : African Academy of Languages, Codice Idee per cultura SRL, E-Africa commission of NEPAD, ENSTA, Funredes, SIL International, Linguasphere Observatory, Language Observatory, International, Multilingual Internet Names Consortium, Organisation Intergouvernementale de la Francophonie (OIF), UNESCO, Unicode IDN in Africa, African Union, ITU. MAAYA has already organized three international symposia on multilingualism on cyberspace[45] and published a reference book on the subject[46] and some of his members have been leaders in the field of measuring languages presence in the web (LOP and FUNREDES).

***Role in SEMACORE***

MAAYA is the main partner in terms of defining the user's requirements, using the research results, coordinating the implication of some members and liaising with international organizations interested in the project outcomes. It has the responsibility of the work-package 2 (Societal issues) and 8 (Dissemination and Exploitation) and is heavily involved with 7 (Uses Cases), managing a set of Uses cases with tangible societal impact and playing a role in products evaluation capitalizing on its previous experience as provider of indicators. MAAYA has been the early designer and promoter of the project, convincing institutional partners (UNESCO, OIF and Union Latine) to fund the first stage of project definition and consortium creation and it will maintain an implicit role of cohesion and synergy provider between the WPs and between the partners as well as a special role of interface with the international organizations motivated by the outcomes of SEMACORE (the already mentioned plus ITU).

***Expected outcome from SEMACORE***

The expected outcomes of SEMACORE in terms of creating indicators and demonstrating the usefulness of linguistic diversity measurement through public policies and economical impacts are totally coherent with the international mission of MAAYA and pave the ground towards the World Summit on Multilingualism which is one current MAAYA objective.

***Key personnel***

**Daniel Prado** is the Executive Secretary of MAAYA. From 1984 to 2011, he was responsible of the Directorate for Terminology and Language Industries (DTIL) of the intergovernmental organization Union Latine. Within the framework of the DTIL, he has promoted the modernization of the romance languages in order to facilitate access to specialized communication in mother tongue. Daniel Prado has coordinated activities related to development of terminologies and language tools, the promotion of the scientific and technical translation and writing, and also the development of linguistic diversity in cyberspace, science, international negotiations, international governance and more. He has been responsible for numerous projects involving national and international spheres in language policy, terminology, language industries and multilingualism in cyberspace, as well as building information systems and specialized multilingual sites (www.portalingua.info, www.terminometro.info, www.hex-libris.info, etc.) and fora. He participated in the creation of several international networks or associations such as Realiter, RITem, Linmiter or EAfT. He coordinates the reference multi-author book "Net.Lang-Challenges of multilingualism in cyberspace", the preparation of the Third International Symposium on Multilingualism in Cyberspace (Paris, November 2012) and activities for the realization of the World Summit on multilingualism.

[45] http://www.maayajo.org/spip.php?rubrique105
[46] http://net-lang.net/lang_en

**Daniel Pimienta** is MAAYA coordinator for Latin America. A French citizen born in Morocco, he read Applied Mathematics in Nice University and holds a Ph.D. in Computer Sciences. After creating a Software House specialized in APL, he joined IBM France (La Gaude Laboratory) and worked 12 years as Telecommunication System Architect and Planner. In 1988, he joined Union Latina, in Santo Domingo, as Scientific Advisor and Head of the REDALC project for creation of LA&C network. In 1993, he launched FUNREDES and focused on ICT4D, defining and managing more than 30 projects with a vision centered on users and contents and a strong research-action component towards proper methodologies. An active civil society player in Information Society themes (representing civil society voice in the World Summit of Information Society - WSIS) with a special perspective on social impact of ICT, virtual communities and linguistic diversity, he is a member of several ICT4D related global groups such as Francophone virtual university, 3EL, GCNP, EUROLATIS, WINDS-LA, REDISTIC, APC, WSIS-AWARD, UN-GAID and Digital Solidarity Fund. He was given, in 2008, the Namur Award (IFIP WG9.2) for his comprehensive actions in the perspective of a positive social impact of ICT. Pimienta is a recognized lecturer and writer on the theme related to Information Society with more than 120 conferences (including 25 keynotes) and 70 publications. Pimienta is a member of the evaluation boards of the Journal of the American Society for Information Science and Technology, the Journal of Community Informatics and the Journal of ICT and Human Development. With FUNREDES, Pimienta has been a leader in the field of measuring language presence in the web since 1998.

**Alvaro Blanco**, a Spanish citizen, graduated in computer science middle degree in La Rioja University, spent the first few years of work experience around computer related tasks: programming, maintenance, user support and trainer. In 2003, he joined FUNREDES taking part actively in projects SOCINFODO and CARDICIS, and especially the "Observatory of the Linguistic and cultural diversity on the Internet", a languages measurement project started by FUNREDES in 1998. He is responsible for development of the integration of machine translation projects in various conferencing systems, from discussion lists to e-learning platforms such as Moodle. Since 2009, he is the Head of the FUNREDES Branch in Spain and work especially in the FUNREDES projects for language measurement in the web.

***Publications***

MAAYA, 2012, Net.Lang : Towards a multilingual cyberspace / Réussir le cyberespace multilingue. Maaya, OIF, UNESCO, Union latine, CRDI, AnLoc. C&F Éditions, 2012.

D. Prado, D. Pimienta, A. Lemoulinier, 2011, Diversité linguistique et cyberespace : état de l'art, enjeux et opportunités, Cosmopolis, N1-4/2011

D. Pimienta, D. Prado & A. Blanco, 2009, Twelve years of measuring linguistic diversity in the Internet: balance and perspectives, UNESCO, CI.2009/WS/1

J. Paolillo, D. Pimienta, D. Prado, 2005, Measuring linguistic diversity on the Internet. A collection of papers edited with an introduction by the UNESCO Institute for Statistics Montreal, Canada.

### B 2.2.3 *Partner 3: Universitat Politecnica de Catalunya, ES (UPC)*

The Technical University of Catalonia (UPC) is a public institution with a vocation to serve society. It offers a broad and higher education in a range of technical, artistic and humanistic fields. UPC is a leading research university in the areas of Architecture and Civil Engineering, Mathematics and Statistics, Social, Human and Life Sciences, Physics and Chemistry, Environment, Energy and Natural Resources, and Information and Communication Technologies. UPC aims at fostering the improvement of educational, research and management activities and greater transparency and accountability to society. The Centre for Speech and Language Applications and Technologies (TALP) at the UPC focuses on research and development in the area human language processing. It is comprised of two research groups: the Speech Processing group and the Natural Language Processing group, and includes over 30 faculty members, 5 full time research personnel and around 40 graduate research assistants. Since its creation in 2002, the centre has been involved in research and development of spoken and written language processing technologies and applications with special attention to English, Catalan, and

Spanish. Activities include construction and use of multilingual lexical resources, question answering, information extraction, machine learning methods, machine translation, text summarization, speech recognition, speech synthesis, natural language interfaces, and core language technologies (such as morpho-syntactic and semantic analyzers, word sense disambiguation, named entity recognition, parsing methods, co-reference resolution, identification of textual entailment, etc.). The group has participated in numerous R&D projects funded by the EU, the Spanish Ministry if Science and Technology (CICyT) and the regional Catalan government. In the last five years, centre members collectivelly published over 200 papers, articles, book chapters and monographs. They organized major international conferences and workshops and served on numerous program committees. In the last ten years, TALP/UPC has participated in several EU projects (NAMIC, FAME, MEANING, LC-STAR, TC-STAR, CHIL, HOPS, FAUST, MOLTO, XLIKE), 10 nationally funded projects, and one DARPA project (Arabic WordNet), all in the area of NLP and Speech processing. Current research focuses on multilingual knowledge acquisition and representation, information extraction, syntactic and semantic parsing, machine learning algorithms, and machine translation.

***Role in SEMACORE***

The UPC TALP research center will participate in SEMACORE providing their expertise in Machine Translation systems and in linguistic analysis tools, in order to develop adaptative translation techniques that will enable the domain and generate adaptation of translation models.

***Expected outcome from SEMACORE***

SEMACORE offers a multilingual and multi-domain environment for applying TALP translation engines and model adaptation methods. Thus, the project poses a challenging set-up that will help to improve and extend these techniques. Also, it offers the chance to extend our open-source language processors to new languages and to improve their coverage and efficiency.

***Key personnel***

**Lluís Padró** is Associate Professor at the UPC since 1999. He obtained his PhD in Artificial Intelligence from UPC in 1998. Prior to becoming professor, he lectured at the UPC from 1991 to 1999. His research ranges across several areas in NLP, with a main focus on basic language processing (tagging, parsing, sense disambiguation, etc) and optimization methods. He has published over 50 papers, many in the most important conferences (ACL, ANLP, NAACL, EACL, COLING, EMNLP, etc.) and major journals (Machine Learning, Computational Linguistics, etc.). He has supervised six PhD theses, and participated in 6 EU funded projects: Acquilex-2, EuroWordNet, NAMIC, MEANING, MOLTO and XLIKE. He also has participated in 10 Spanish government funded projects, acting as local coordinator in 4 of them. He is the main leader of the FreeLing project, an open-source library of natural language analysis methods for multiple languages.

**Lluís Màrquez** is Associate Professor at the Technical University of Catalonia (UPC) since 2000. PhD in Computer Science (UPC 1999; awarded the UPC prize for doctoral dissertations in Computer Science). His research focuses on Machine Learning methods for Natural Language structure prediction problems, including syntactic and semantic parsing, and statistical machine translation. He has 100+ papers in Natural Language Processing and Machine Learning journals and conferences. He has been Program Chair and Area Chair of major conferences in the area, including ACL, EACL, EMNLP, CoNLL, EAMT, etc. and organized several international evaluation tasks at Senseval/ SemEval (2004, 2007, 2010) and CoNLL shared tasks (2004-2005, 2008-2009). Secretary and President of the ACL SIG on Natural Language Learning (SIGNLL)

***Publications***

Ramona Enache and Cristina España-Bonet and Aarne Ranta and Lluís Màrquez: A Hybrid System for Patent Translation, Proceedings of the 16th Annual Conference of the European Association for Machine Translation pg. 269--276. European Association for Machine Translation. May, 2012.

Cristina España-Bonet and Lluís Màrquez: Robust Estimation of Feature Weights in Statistical Machine Translation, Proceedings of the 14th Annual Conference of the European Association for Machine Translation pg. 190--197. May, 2010.

Jesús Giménez and Lluís Màrquez: On the Robustness of Syntactic and Semantic Features for Automatic MT Evaluation, Proceedings of the Fourth Workshop on Statistical Machine Translation pg. 250--258. Association for Computational Linguistics. Association for Computational Linguistics. March, 2009.

Xavier Carreras AND Michael Collins: Non-projective Parsing for Statistical Machine Translation, In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP), 200-209, Singapore, 2009.

### B 2.2.4 *Partner 4: DIALOGIC, NL (DIALOGIC)*

Dialogic is an independent research-based consultancy firm located in Utrecht, the Netherlands. Dialogic focuses on processes of innovation IT, telecom, new media developments, and services. Dialogic employees have been active in these fields for a long time and have performed many projects in the area of telecommunications and media studies, conditions for introducing new technologies, assessments of user requirements, technology-based foresight studies, scenario-studies and economic industry analysis. In the area of telecommunications and media, Dialogic has developed towards an established centre of expertise, both covering the Netherlands and increasingly Europe. In particular, comparative and benchmarking studies have been performed, in the areas of broadband internet penetration and user pattern development, and the development of user needs. With the internet as data source project, Dialogic has introduced the automated collection of data from the internet for statistical purposes (including the notion of user-centric measurements).[47] The initial project (2006/2007) was geared towards the Dutch Bureau of Statistics.[48] Several follow-up projects have recently been completed (EC, OECD) or are planned for 2013[49].

***Role in SEMACORE***

Dialogic will offer its specific expertise on internet-based data collection models and will develop and built the user-centric measurement client for language detection in actual use. Dialogic will also provide parts of the overall quality management of the project (audit external validity of the research results).

***Expected outcome from SEMACORE***

The application of user-centric measurements at this level of detail is a new and unique development track. Moreover, the combination of the client with a survey model that can be triggered by the data from the client renders it into a very powerful panel survey instrument. The language component is entirely new for Dialogic and is an exciting new area to enter.

***Key Personnel***

**Robbin te Velde** is principal researcher at Dialogic. He has extensive experience in ICT research projects. During the last 20 years he has been alternatively working at technical universities (Twente University, Delft University of Technology, Eindhoven University of Technology, research consultancies, and think

[47] See http://ec.europa.eu/information_society/newsroom/cf/itemdetail.cfm?item_id=8701

[48] http://www.rijksoverheid.nl/documenten-en-publicaties/rapporten/2008/05/13/go-with-the-dataflow-main-report%5B2%5D.html

[49] Reg Brennenraedts, Robbin te Velde, Tommy van der Vorst, Arthur Vankan (2012): Estimating Online Advertisement Expenditure Using Third Party Data: A feasibility study. Report for the OECD, Utrecht: Dialogic. Presented at the Working Party on Indicators for the Information Society, Paris, December 14 2012.

tanks (including the Rand Corporation). He has a strong background in methodology and is specialized in international comparative studies. Next to strategic IT consulting he has implemented hands-on IT-projects for large telecom operators (including BT and KPN) and several multinationals. Besides completing over 100 research projects he has written a large number of scientific articles on a wide range of areas such as international politics, philosophy, knowledge management, business administration, technology policy and information management. Robbin has been principal researcher in all Dialogic projects with regard to automated data collection.

**Reg Brennenraedts** is partner & senior researcher at Dialogic. He holds a bachelor in Electrical Engineering, a Master of Science in Innovation Science (Eindhoven University of Technology) and an MBA in strategy and corporate finance (TiasNimbas Business School). He has been mainly working in the fields of telecom and IT. He is a regular advisor to major actors in the telecom domain and Dutch Ministry of Economic Affairs. With regard to IT he has been involved in innovation projects with regard to gaming, digital music, digital radio (DAB) and digital TV. Reg is a seasoned project leader and has also been in charge of all automated data collection projects at Dialogic.

***Publications***

Reg Brennenraedts, Robbin te Velde (2012): Feasibility Study on Statistical Methods on Internet as a Source of Data Gathering. Report for the European Commission, DG Communications Networks, Content & Technology

Robbin te Velde, Reg Brennenraedts (2011): Measuring the Impact of ICT on Health Care. In: R. Weiss (ed.) The Linked World: How ICT is Transforming Societies, Cultures and Economies. New York: Conference Board (Ch.6).

Robbin te Velde (2009): Public Sector Information: Why Bother? In: P. Uhlir (ed.) The Socioeconomic Effects of Public Sector Information on Digital Networks. Towards a better understanding of different access and reuse policies. Washington DC: National Academies Press (Ch.6)

Rex Arendsen, Robbin te Velde, Tom Engbers (2006): An Empirical Study on Business-to-Government Data Exchange Strategies to Reduce the Administrative Costs for Businesses. In: R. Suomi et al. (eds.). IFIP International Federation for Information Processing. Volume 226, Project E-Society: Building Bricks, pp. 311-323.

### B 2.2.5 *Partner 5: Centre National de la Recherche Scientifique, FR (CNRS)*

*Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (LIMSI)*

Founded in 1939, the CNRS (National Center for Scientific Research) is a government-funded research organization, under the administrative authority of France's Ministry of Research. (http://www.cnrs.fr). The CNRS encourages collaboration between specialists from different disciplines and has laboratories located throughout France. LIMSI is one of France's largest research laboratories working on language technologies; it covers the full spectrum from low level signal processing to spoken and written language processing and machine translation. The main activities of the Spoken Language Processing Group cover the following domains: Speech Recognition, Speech Understanding, Dialog Systems, Speaker and Language Recognition, Speech Translation and Audio Indexation. Associated activities include data collection, system evaluation and technology transfer. The group has succeeded in basic research as well as in applied research developing new algorithms, prototypes and databases. Advanced commercialized products developed from studies at LIMSI are now being used in several applications. LIMSI has participated in a number of projects on speech recognition (CORETEX), spoken language systems for information retrieval (EC MASK, Railtel, Arise, Home, DISC and Amities) and audio document indexation and retrieval in multiple languages (EC Olive, Alert, Echo, RNRT Theoreme, AudioSurf), facilitating human-human communication (CHIL) and spoken machine translation (TC-STAR). LIMSI, in collaboration with Vocapia Research, has developed competitive transcription systems for broadcast data in 9 languages in the context of the Quaero program. (web site: http://www.limsi.fr/tlp)

***Role in SEMACORE***

The main contributions from CNRS-LIMSI to SEMACORE will be in the development of robust statistical models for language identification, content extraction and opinion analysis, and unsupervised learning for multilingual speech recognition valid for the wide variety of audio data types found on the web.

***Expected outcome from SEMACORE***

The CNRS will improve their models and technology for language recognition and speech analytics in heterogeneous data for a large number of languages. SEMACORE will also provide a means to better access to language resources in many languages.

***Key personnel***

**Lori Lamel** is a senior CNRS researcher in the Spoken Language Processing group at LIMSI which she joined in October 1991 She received her PhD degree in EECS in May 1988 from the Massachusetts Institute of Technology. Her principal research activities are in speech recognition; acoustic-phonetic studies; lexical and phonological modeling; and conversational systems. She has been a prime contributor to the LIMSI participations in speech recognizer benchmark evaluations and developed the American English pronunciation lexicon. She has been involved in many European projects, most recently leading the LIMSI activities in the IP Chil and the Speech Processing activities in Quaero. Dr. Lamel is a member of the Speech Communication Editorial Board, was a member of the Interspeech International Advisory Council, the IEEE James L. Flanagan Speech and Audio Processing Award Committee (2006-2009) and the EU-NSF Working Group for 'Spoken-Word Digital Audio Collections'. She has over 250 reviewed publications and is co-recipient of the 2004 ISCA Best Paper Award for a paper in the Speech Communication Journal.

**Jean-Luc Gauvain** is a senior scientist at the CNRS and head of the Spoken Language Processing Group at LIMSI-CNRS. He received a doctorate in Electronics from the University of Paris XI in 1982, and has been a permanent CNRS researcher at LIMSI since 1983. His primary research centres on large vocabulary speech recognition, language identification and audio indexing. His research interests also include conversational interfaces, speaker recognition, and speech translation. He has participated in many speech-related projects both at the French National and European levels and has led the LIMSI participation in numerous DARPA/NIST organized evaluations since 1992. He has over 280 publications and received the 1996 IEEE SPS Best Paper Award in Speech Processing and the 2004 ISCA Best Paper Award for a paper in the Speech Communication Journal. He was co editor-in-chief of the Speech Communication from 2007 to 2009. He was awarded a CNRS silver medal in 2007, and was an appointed member of the National Committee of Scientific Research from 2008 to 2010. Since April 2008, he is the Scientific Coordinator for the Quaero programme.

***Publications***

L. Lamel and J.L. Gauvain. Speech recognition. In R. Mitkov, editor, OUP Handbook on Computational Linguistics, chapter 16, pages 305322. Oxford University Press, 2003.

J.L. Gauvain, L. Lamel, and G. Adda. The LIMSI Broadcast News Transcription System. Speech Communication, 37(1-2):89108, 2002.

M. Faouzi BenZeghiba, J.L. Gauvain and L. Lamel, Gaussian Backend Design for Open-set Language Detection, IEEE ICASSP'09, Taipei, April 2009

L. Devillers, L. Vidrascu, and L. Lamel. Challenges in real-life emotion annotation and machine learning based detection. Journal of Neural Networks, 18/4, 2005.

### B 2.2.6 *Partner 6: KYOS IT Security, CH (KYOS)*

Based in Switzerland, Kyos is a service-related company specialised in the IT security field. Kyos offers design and project management for security related solutions. In its engineering activities, Kyos provides high qualified consultants for projects like vulnerability assessment and testing, system integration as well as network and application security solutions. Maintaining the security level related to its' investment is a complex task of administration, analysis and permanent monitoring. In this field, Kyos proposes an offer of customised services like security equipment administration & supervision, personalised security threats, monitoring, periodic vulnerability assessment and log management & analysis. Kyos was involved in SEINIT and DEMONS projects with a strong focus in the trust management, privacy, intrusion detection and honeypot field. To test the level of protection of a company against trojan horse & malicious code, Kyos implements a testing framework (NOSE) using multiple diffusion, installation & information retrieval methods. Thanks to our research activity, this framework is constantly updated according to the evolution of content filtering and host based security solutions. Kyos has conducted several projects in the field of security audit, for example for an e-voting platform and also for e-banking solutions. Therefore, Kyos has developed a strong know-how in privacy vulnerabilities and privacy protection. (web site: http://www.kyos.ch)

***Role in SEMACORE***

Kyos main activity will concern WP2 Societal Issues. Kyos will also address any intellectual property, legal, regulatory or ethical issues raised by the research conducted in the project in coordination with all partners. Kyos will also be in charge of defining privacy protection and security technologies architecture. Kyos auditors will also participate in WP4 by performing an audit of the user-centric measurement system in order to identify security and privacy vulnerabilities.

***Expected outcome from SEMACORE***

The output of SEMACORE will improve the spectrum of the company's security consultancy services. Everyday KYOS helps its customers to define, integrate and maintain the best solutions for their security. Currently we observe more and more demand of advice & consultancy related to privacy, in many "classical" disciplines like: Identity Management and Access Control, Intrusion detection, Log Analysis & Correlation etc. Thanks to SEMACORE the knowledge gained in this area will give KYOS additional credits and expertise to give state of the art technology consultancy services to our clients.

***Key personnel***

**Fabien Jacquier** is a telecommunication and information technology engineer. His experiences in network and security analysis and implementation give him a wide approach of security concepts, especially in the field of privacy and security assessment. He has participated to several FP6 and FP7 projects as a research engineer in the field of IPv6, Next Generation Networks, and security. Co-founder and director of Kyos since 2003, he also works as a security consultant for Kyos customers.

**Maxime Feroul** has an engineering degree (97) in Telecommunication and Information technology. His working experience is in security consulting and complex internet hosting. He was responsible to design and implement several security architecture and drive vulnerability assessment for customers with a high level security needs such as bank and nuclear related industry. He gained a deep experience in the field of authentication, identity management and federation, vulnerability assessment and intrusion detection. Co-founder and technical director of Kyos since 2003, he works now as a security consultant for Kyos customers and as a research engineer in the field of security. He is currently involved in projects dealing with j2ee and web services security layer.

**Eric Lederrey** has an engineering degree in information technology. He is a security expert and works as an auditor and ethical hacker within Kyos team for three years. He has conducted a research project named Auditbay, which has for objective to establish an information system for the implementation / audit trail in order to centralize all the information necessary to conduct an audit: official documentation, script and internal revenue, project documentation, generate disk images ready to be

deployed, generation of customer documentation. His recent experience of e-voting code audit included many privacy issues.

### B 2.2.7 *Partner7: IAIS/Fraunhofer, DE (FRAUNHOFER)*

The Fraunhofer-Gesellschaft zur Förderung der angewandten Forschung e.V. (Fraunhofer Society for the Advancement of Applied Research (FRAUNHOFER)) is Germany's leading organisation of research institutions for applied research. Fraunhofer's 60 research institutes are co-ordinating and/or contributing to large national and international industrial applied research projects, as well as to research projects targeting the service sector, national and regional governments and the EU. A staff of 18 000, the majority of whom are scientists and engineers (with university degree), generate the annual research budget of more than €1.6 billion.

The Knowledge Discovery Department (KD) at the Fraunhofer Institute for Intelligent Analysis and Information Systems (IAIS) in Sankt Augustin, Germany, is a research group (50 scientists) located in the field Machine Learning and Data Mining. Professional experience and expertise in this group include Data Mining, Text Mining, Statistical Relational Learning, Geographic Information Systems, and Distributed Computing. The group has extensive experiences with EU projects, having coordinated in the last 6 years five ICT projects and participated in more than 18. To mention a few, FRAUNHOFER coordinates the LIFT project (local inference in massively distributed systems, 2010-2013), the AntiPhish project (Anticipatory Learning for Reliable Phishing Prevention, 2006-2009) and the KDubiq Coordination Action (Knowledge Discovery in Ubiquitous Environments, 2006-2008), and is involved in the ACGT project (Advancing Clinico-Genomic Trials on Cancer, 2006 - 2009), the LifeWatch project (development of an European infrastructure for coping with the biological diversity, 2007 – 2009), the SCY project (Science Created by You, 2008-2012), the DICODE project (Mastering Data-Intensive Collaboration and Decision Making, 2010-2012), the INSIGHT project ( Intelligent Synthesis and Real-time Response using Massive Streaming of Heterogeneous Data), and the IP SIMDAT (Grids for Industrial Product Development, 2004-2008). The group is also currently involved in several nationally funded and industrial projects in Multimedia Mining, Spatial Data Mining, and Text Mining.

***Role in SEMACORE***

Fraunhofer will provide expertise in multilingual content analysis (detection of content categories and opinions on concepts). Fraunhofer will also coordinate System Development and Integration of SEMACORE and provide a large Big Data infrastructure.

***Expected outcome from SEMACORE***

Fraunhofer will extend its expertise on content recognition to a Big Data environment. Especially relevant is the system architecture and system development in this case. We envision many and very relevant public and commercial applications of multi-lingual opinion mining on Big Data.

***Key personnel***

**Gerhard Paass** studied mathematics and computer science and received a PhD in economics at the University of Bonn. Dr. Paass led the EU project DIASTASIS on text classification for web mining. He was co-ordinator of the EU project AntiPhish (Anticipatory Learning for Reliable Phishing Prevention, 2006-2009) for filtering phishing emails. He also led several industry projects, e.g. the project MediaRank for the assessment of commercial text classification systems and a project to classify fraudulent eBay offers. He organized a number of workshops on multimedia learning, ontology learning, and text mining for security at international conferences and served on the program committee for several international conferences and journals like such as ECML-PKDD, UAI, KDD, ICDM, SDM, SIGIR, MLJ, and DAMI. Currently, he leads a subproject of the German joint semantic THESEUS program (New Technologies for the Internet of Services, 2006-2012) on semantic technologies.

**Karl-Heinz Sylla** studied Mathematics at the University of Cologne. He works at Fraunhofer IAIS as project leader and system architect. His interests are topics of software engineering. He was consultant of industrial projects and gave many internal and external courses on system architecture, software design and agile development methods. At Fraunhofer IAIS he applied these topics to big customer projects, for example a planning system for marketing purposes at Deutsche Post based on complex and big data sets about target group qualities. In 2012 he was responsible for design and construction of a Big Data application that combines state of the art technologies for web crawling, massive data storage and analytical tasks for batch and real-time processing.

**Michael May** obtained his PhD from the graduate program in Cognitive Science, Univ. Hamburg, working on machine learning of causal relationships. He is head of Fraunhofer IAIS Knowledge Discovery department and leads the research efforts at the intersection of data mining, machine learning, and spatial technologies. His current main research interest is application of data mining to structural data. He was and is coordinator of several European projects, including the FET-Open LIFT (Local Inference in Massively Distributed Systems, 2010-2013), FET-Open KDubiq Knowledge Discovery in Ubiquitous Environments Coordination Action (2005-2008) and the KDNet Knowledge Discovery Network of Excellence (2002-2005). He has been local chair of the International Conference on Machine Learning ICML 2005, of ILP 2005, and chairman of the working group Data Management in the FP6 EU Grid Concertation Forum from 2004 to 2006. Dr. May has been principal investigator in several recent industry funded projects on spatial learning.

***Publications***

M. Neumann, B. Ahmadi, K. Kersting (2011): Markov Logic Sets: Towards Lifted Information Retrieval Using PageRank and Label Propagation. Proc. Conference on Artificial Intelligence AAAI, San Francisco 2011.

Gerhard Paass, Frank Reichartz (2009): Exploiting semantic constraints for estimating supersenses with CRFs. Proc. International Conference on Data Mining (SDM), Spraks, Nevada 2009.

Anja Pilz, Gerhard Paass (2012): Collective Search for Concept Disambiguation. International Conference on Computational Linguistics, Coling 2012, Mumbai, India.

Frank Reichartz, Hannes Korte, Gerhard Paass (2010): Semantic relation extraction with kernels over typed dependency trees. Proc. ACM SIGKDD Conference on Knowledge, Discovery, and Data Mining, KDD 2010, Washinghton DC.

### B 2.2.8 *Partner 8: Stichting Centrum voor Wiskunde en Informatica, NL (CWI)*

The Stichting Centrum voor Wiskunde en Informatica (CWI) is the Dutch national research institute for mathematics and computer science. It is a private, non-profit organization located at the Science Park Amsterdam. CWI's mission is twofold: To perform frontier research in mathematics and computer science, and to transfer new knowledge in these fields to society. CWI actively pursues joint projects with external partners, provides consulting services, and stimulates the creation of spinoff companies. CWI also manages the Benelux Office of the W3C and hosts both the Semantic Web Activity Lead and the chair of the XHTML and XForms Working Group. CWI is strongly embedded in Dutch university research: about twenty-five of its permanent senior researchers hold part-time positions as professors at universities and many projects are carried out in cooperation with university research groups. CWI receives a basic funding from the Netherlands Organisation for Scientific Research (NWO), amounting to about two third of the institute's total income. The remaining third is obtained through national research programmes, international programmes, and contract research commissioned by industry. CWI hosts a staff of 235 full time employees, 50 permanent scientific staff, 135 temporary scientific staff, and 50 support staff.

***Role in SEMACORE***

Centrum Wiskunde & Informatica brings in their expertise on data management and information retrieval, in particular a long track record of research and applications where the two areas meet. In the VITALAS (FP6 IP) project for example, the XML information retrieval system developed in this group performed the key search operations, including the analysis of user interaction logs for the suggestion of keywords and (multimedia) concepts. The objective of a declarative language for defining and manipulating language models over large amounts of data fits perfectly with the long term focus on the integration of databases and information retrieval. Ongoing research in the Interactive Information Access group (INS2) addresses more related topics, including user search log analysis, entity retrieval and social media, focusing on how structured information acquired from linked open data can augment interactive information access to semi-structured text and multimedia collections.

***Expected outcome from SEMACORE***

CWI will deepen its experience with language modelling approaches, in a much more diverse and heterogeneous setting than the cultural heritage and enterprise data we have worked with before. The methods for inducing topic maps to summarize large amounts of web data will help progress in our research programme where we mix statistical and logical models to further information retrieval. We finally expect to exploit the gained expertise in handling large amounts of multi-lingual web data together with spin-off company Spinque.

***Key personnel***

**Arjen P. de Vries** is a tenured researcher at CWI leading the Interactive Information Retrieval research group, and a full professor (0.2 fte) in the area of multimedia data management at the Technical University of Delft. De Vries received his PhD in Computer Science from the University of Twente in 1999, on the integration of (multimedia) information retrieval and database systems. He is especially interested in the design of database systems that support search in multimedia digital libraries. He has worked on a variety of research topics, including (multimedia) information retrieval, database architecture, query processing, retrieval system evaluation, and ambient intelligence. In recent research, he concentrates on the question how to exploit the traces left by people online as links, clicks, and participation in social networks, to improve the quality of information search. He has participated in EU projects VITALAS and PuppyIR, Dutch national programmes MultimediaN and the new COMMIT, and is a member of Cost Action MUMIA and the EU PetaMedia Network of Excellence. He has supervised several best student papers at ACM conferences. He has been general co-chair of the ACM SIGIR 2007 conference in Amsterdam and programme co-chair of CIKM 2011 and ECIR 2012. He coordinates international benchmarking activities at TREC (ranking entities on the WWW). De Vries is a member of the TREC PC, and a steering committee member of INEX (the Initiative for the Evaluation of XML Retrieval).

**Jacco van Ossenbruggen** is affiliated with the Interactive Information Access group at the Centrum voor Wiskunde en Informatica (CWI) in Amsterdam, and with the web & media research group at VU University in Amsterdam. His research interests include user interfaces for unreliable data, web-based metadata modeling and integration, and data provenance on the web. He is currently researching these topics in the cultural heritage domain (as part of the European PrestoPrime and the Dutch national COMMIT projects) and in the marine biology domain (in the European Fish4Knowledge project) . He obtained a PhD in computer science from VU University Amsterdam in 2001.

***Publications***

Arjen P. de Vries et al. (2008): Overview of the INEX 2007 entity ranking track. Focused Access to XML Documents, 6th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2007 Dagstuhl Castle, Germany.

M. Hildebrand, J. van Ossenbruggen (2012): Linking User-Generated Video Annotations To The Web Of Data. In: Proceedings of 18th International Conference on Multimedia Modeling 2012.

Jiyin He, Vera Hollink, Arjen P. de Vries (2012): Combining Implicit and Explicit Topic Representations for Result Diversification. In Proc. of the 35th international ACM SIGIR conference on Research and development in information retrieval, SIGIR'12, Portland, USA.

Karl Gyllstrom, Carsten Eickhoff, Arjen P. de Vries, Marie-Francine Moens (2012): The Downside of Markup: Examining the Harmful Effects of CSS and Javascript on Indexing Today's Web. Proc. ACM International conference on Information and knowledge management, CIKM'12, Maui, HI, USA.

### B 2.2.9 *Partner 9: Vocapia Research, FR (VOCAPIA)*

Founded in 2000, Vocapia Research is an R&D company specialized in the development of multilingual technologies for speech and language processing, in particular speech to text transcription systems, audio and speaker segmentation and identification, and language recognition. It has privileged partnerships with the CNRS-LIMSI laboratory.

Statistical methods are used in the VoxSigma software suite to model spoken language and to build leading edge speech processing technologies which can serve a variety of applications, in particular for automatic audio indexing and speech analytics. Large vocabulary continuous speech recognition is a key technology that can be used to enable content-based information access in audio and video documents since most of the linguistic information is encoded in the audio channel of audiovisual data, which once transcribed can be accessed using text-based tools. Via language identification, speech recognition, and speaker recognition, spoken document retrieval can support random access using specific criteria to relevant portions of audio documents, reducing the time needed to identify recordings in large multimedia databases. Vocapia Research's VoxSigma speech-to-text systems cover many languages.

Vocapia Research and CNRS-LIMSI technologies have been ranked first in the French Technolangue ASR benchmark tests (2005 and 2009) and in the Dutch NBEST ASR benchmarks test (2008). Vocapia Research is providing and further developing these technologies for the Quaero program: the Exalead video search engine has integrated the CNRS-LIMSI/Vocapia text-to-speech technology.

Vocapia Research participates in the IARPA program Babel as part of the Babelon team using the speech-to-text output in order to provide effective search capability for analysts to efficiently process massive amounts of real-world recorded speech.

***Role in SEMACORE***

Vocapia will offer expertise in speech to text transcription systems, language recognition, audio and speaker segmentation and identification.

***Expected outcome from SEMACORE***

Vocapia recognizes the importance of adapting their technologies to deal with the enormous quantity of heterogeneous data types found on the web. SEMACORE will enable Vocapia to improve their technologies, test these in real-world condition and extend their language offer.

***Key personnel***

**Viet-Bac Le** is a research scientist at Vocapia Research which he joined in October 2010. He received his PhD degree in Computer Science from the Joseph Fourier University and the CLIPS-IMAG Laboratory, Grenoble in 2006. From 2006 to 2008, he was Postdoctoral Fellow at the LORIA laboratory (Nancy) and at LIG laboratory (Grenoble). He joined the Spoken Language Processing Group at LIMSI-CNRS, Orsay as research associate from 2008 to 2010. Dr. Le has been involved in several speech-related French National projects, the Quaero program, the DARPA GALE program and the IARPA Babel program. His research interests include speech recognition, audio segmentation, speaker identification, speech translation and keyword spotting. He has over 30 reviewed publications.

**Bianca Vieru** joined Vocapia Research as research scientist after obtaining her PhD degree in Computer Science from Paris XI University and the LIMSI laboratory in 2008. Her research focused on the characterization and the identification of foreign accents in French. She received a Master in Cognitive Science in 2004 from Paris XI University and a BSc in Computer Science in 2001 from Bordeaux I University. She has several reviewed publications in conference and journals. Dr. Vieru has been involved in several speech-related French National projects, the Quaero program and the IARPA Babel program. She is currently working on the development of speech-to-text systems in multiple languages.

**Cecile Woehrling** obtained her PhD degree in Computer Science from Paris XI University and the LIMSI laboratory in 2009 after which she joined Vocapia Research as research scientist. Her research focused on characterizing and identifying regional French accents. She received a Master in Cognitive Science in 2005 from Paris XI University and a BSc in Computer Science in 2004 from Paris XI University. She has published several conference and journals papers on speech. Dr. Woehrling has been involved in several speech-related French National projects, the Quaero program and the IARPA Babel program. At Vocapia Research she works on developing acoustic, linguistic and pronunciation models for speech-to-text systems.

**Publications**

L. Lamel and B. Vieru. Development of a Speech-to-text transcription system for Finnish. In The second International Workshop on Spoken Languages Technologies for Under-resourced languages(SLTU10), pages 62?67, Penang, Malaysia, May 2010.

L. Lamel, S. Courcinous, J. Despres, J-L. Gauvain, Y. Josse, B. Vieru, C. Woehrling et al. Speech Recognition for Machine Translation in Quaero. IWSLT, San Francisco, CA, USA, 2011.

J. Despres, P. Fousek, J-L. Gauvain, S. Courcinous, Y. Josse, L. Lamel, A. Messaoudi. Modeling Northern and Southern Varieties of Dutch for STT.Interspeech'09, 96-99, Brighton, UK, 2009

J. Despres, L. Lamel, J.L. Gauvain, B. Vieru, C. Woehrling, V.B. Le, I. Oparin The Vocapia Research ASR Systems for Evalita 2011. Book Section, Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2013.

### B 2.2.10 *Partner 10: Nielsen (NIELSEN)*

**The Nielsen Company** is a leading global information and measurement company that provides clients with a comprehensive understanding of consumers and consumer behaviour. Nielsen deliver's critical media and marketing information, analytics and industry expertise about what consumers watch (consumer interaction with television, online and mobile) and what consumers buy on a global and local basis. Our information, insights and solutions help our clients maintain and strengthen their market positions and identify opportunities for profitable growth. We have a presence in approximately 100 countries, including many emerging markets. We hold market leading positions in many of geographies. Based on the strength of the Nielsen brand, our scale and the breadth and depth of our solutions, we believe The Nielsen Company is the global leader in measuring and analysing consumer behaviour in the segments in which we operate.

We help our clients enhance their interactions with consumers via marketing and make critical business decisions that we believe positively affect our clients' sales. Our data and analytics solutions, which have been developed through substantial investment over many decades, are deeply embedded into our clients' workflow as demonstrated by our long-term client relationships, multi-year contracts and high contract renewal rates. The average length of relationship with our top ten clients, which include The Coca-Cola Company, NBC Universal, Nestle S.A., News Corp., The Procter & Gamble Company and the Unilever Group, is more than 30 years. Typically, before the start of each year, nearly 70% of our annual revenue has been committed under contracts in our combined Watch and Buy segments.

We align our business into two reporting segments, the principal two of which are **What Consumers Watch** (media audience measurement and analytics) and **What Consumers Buy** (consumer purchasing measurement and analytics). Our Watch and Buy segments, which together generated 96% of our revenues in 2009, are built on an extensive foundation of proprietary data assets designed to yield essential insights for our clients to successfully measure, analyse and grow their businesses. The information from our Watch and Buy segments, when brought together, can deliver powerful insights into the effectiveness of advertising by linking media consumption trends with consumer purchasing data to better understand how media exposure drives purchase behaviour. We believe these integrated insights will better enable our clients to enhance the return on investment of their advertising and marketing spending.

***Key personnel***

**René Lamsfuß** is Vice President Market Governance & Data Strategy Europe, at Nielsen Europe – leading global provider of information and analytics around what consumers watch and buy. He was until 2011 in charge of Nielsen's syndicated digital product set in Europe before he has taken over his new role. Mr Lamsfuss ensures strong working relationships with industry bodies, joint industry bodies and key clients in order to deliver excellent products that fully comply to privacy and data protection guidelines on both, national and regional level. Mr Lamsfuss joined Nielsen Online from United Internet Media, the leading German Online Sales House, where he was responsible for market research and media consulting as well as for managing all relations with industry bodies and joint industry committees. In his position at United Internet Media, he was also the Architect of a number of innovative fusion and product development opportunities, such as Targeting. In 2008 and 2009, Mr Lamsfuss was appointed board member of IAB Europe. He furthermore served as Chairman of AGOF, the German industry body for Online Media Research, where he managed – amongst others - all activities related to privacy and data protection regulation. In September 2011, Mr Lamsfuss was appointed as Chairman of the IAB Europe Online Research Committee, which works together with IAB Europe and local IABs as well as members such as Google, Adobe or ComScore on market governance for online research, especially vis-à-vis decision-makers in Europe. René Lamsfuss holds a MA degree in Geography, Political and Social Science from Heinrich-Heine-University in Düsseldorf.

***Publications***

Nielsen/NMIncite (2012): "State of the Media. The Social Media Report."
Nielsen/NMIncite (2012): "The Customer-First Imperative. 5 Steps for Applying Social Media to Generate Transformative Consumer Insights."
Nielsen/NMIncite (2012): "The Social Care Imperative. Four Steps to Drive Brand Health and Customer Acquisition."
Nielsen/NMIncite (2012): "Insights Throughout the CPG Brand Lifecycle."

## B 2.3 Consortium as a whole

### B 2.3.1 Consortium overview and role of the participants

The selected group of partners participating in SEMACORE project is highly capable to conduct the tasks associated to this project, as it has been shown, due to their deep involvement in European and International infrastructures and European Technology Platforms and Technical research capacities.
All of them have been selected because of their own personal and professional reputation on the field of support to R&D computer research expertise, business & linguistic expertise, Internet, ICT Components and systems. Therefore, recommendations from SEMACORE partners have a high potential

to be accepted and adopted to strengthening the technical capacities for supercomputing in Europe guaranteeing the biggest possible impact at worldwide level.
SEMACORE partners are fully convinced that their participation in this project will significantly contribute to classification of metadata and use supercomputing methods not only for industrial data interests, but on all aspects of the internet including linguistic indicators by identifying strategic partners and by developing international policy objectives and market development priorities, as well as providing a high level competence guidance to assist discussions setting up privilege partnerships.
The SEMACORE partners are excellently complementing each other covering the necessary perspectives to deal with the most relevant areas of the work programme: Technical experts on metadata analysis and research, business consultancy expertise, ICT industry associations and socioeconomic consultants, as well as industrial & linguistic experts for the applications on pilots.

The consortium represents a well balanced partnership conceived to reach following objectives:

- Collaborative research activities for building indicators for spaces other than the traditional web and for other approaches than those involving a static vision of existing resources.
- Combination of various research and metadata exploitation methods and analysis to create relevant indicators for complex measurements on the ICT systems
- Application of results on real sectors (business and language policies) through the Uses Cases

All the partners are well-established in their Countries and cover different fields of knowledge. The integration of these peculiarities will ensure the highest level of commitment.
All partners have the necessary competences and in-house resources to carry out the activities planned.

### B 2.3.2 Complementarity of participants

Underneath is a table outlining the complementarities of the consortium partners.

| Role | Partner | Contribution |
|---|---|---|
| Coordinator | GEIE ERCIM | Project management, financial & administrative coordination |
| Research Partners | CNRS<br>UPC<br>FRAUNHOFER<br><br>CWI | Analysis of spoken language in multimedia data<br>Crawling, sampling, network link analysis<br>content analysis, system development & integration, scientific coordination<br>Data & knowledge representation |
| Industrial Partners | DIALOGIC<br>KYOS<br>VOCAPIA<br>NIELSEN | Data collection, user centric measurement, quality assurance<br>Privacy protection and security technologies architecture<br>Speech to text, audio segmentation, language recognition<br>Applications and pilots |
| Institutional Partner | MAAYA | Societal issues, Dissemination & exploitation<br>Result assessment and evaluation<br>Policy oriented Uses Cases |

**Figure 16 - Complementarity of consortium partners**

### B 2.3.3 Sub-contracting

None.

### B 2.3.4 Other countries

None.

### B 2.3.5 Additional partners

The SEMACORE consortium is complete at the time of submitting the proposal. No other partner is foreseen or needed in the foreseeable future.

## B 2.4 Resources to be committed

### B 2.4.1 Overview of resources to be committed

SEMACORE resources have been carefully planned. The project realisation requires mainly high level human resources. The overall planned effort is **354 person-months**, or close to 30 person-years, representing a global eligible cost of personnel in excess of 3.5 M€. The project duration is 36 months. This means that SEMACORE will act as a geographically distributed research team of 20+ high level scientists and engineers. The figure below shows the effort allocation across the 8 SEMACORE work packages, showing that most resources are distributed between the RTD work packages (yellow).

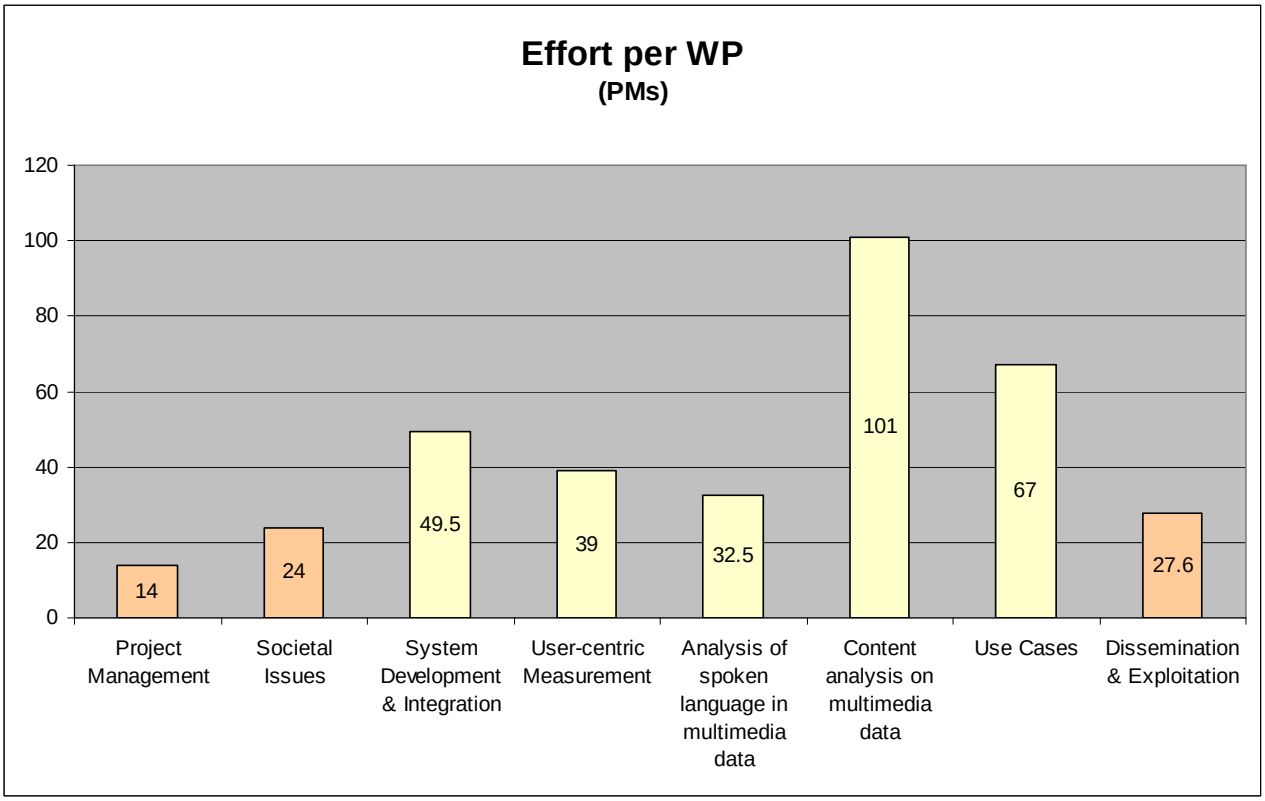


**Figure 17 - SEMACORE effort (PM) allocation per work package**

The overall cost structure breakdown into research activities (RTD), management (MGT), and OTHER activities such as dissemination, exploitation and collaboration is presented in the figure below.

Management represent 5.6% of the total eligible costs and less than 7% of the total requested EC contribution. The **requested EC contribution is 2 998 338 €** representing 75% of the estimated eligible costs.

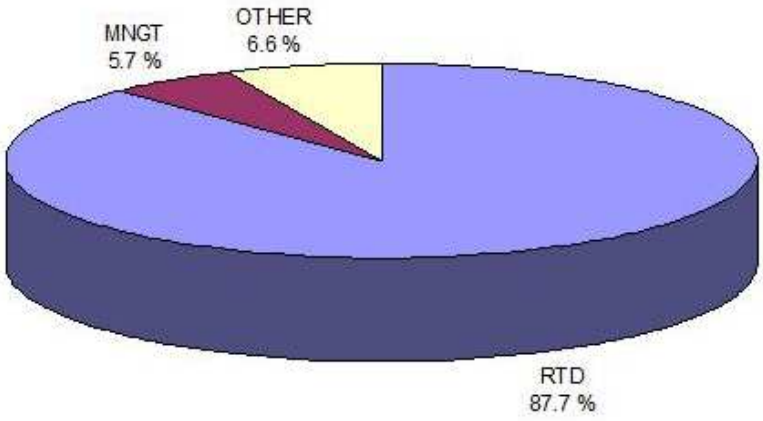


**Figure 18 - SEMACORE breakdown by cost categories**

The table below represents the consortium budget as it has been entered in the proposal submission system. Further explanations are proposed in the following paragraphs.

## A3.2: Budget

### Estimated budget in EUR (whole of the project)

| Nr. | Organisation Short Name | Organisation country | RTD | Demonstration | Management | Total | Total receipts | Requested EU contributions |
|---|---|---|---|---|---|---|---|---|
| 1 | ERCIM | France | | | 285 235 | 285 235 | 0 | 285 235 |
| 2 | Maaya | CH | 390 825 | 0 | 124 875 | 515 700 | 0 | 417 993 |
| 3 | UNIVERSITAT POLITECNICA DE | ES | 195 253 | | 2 515 | 197 768 | 0 | 148 954 |
| 4 | DIALOGIC INNOVATIE & INTERACTIE BV | NL | 494 550 | 0 | 6 937 | 501 487 | 0 | 377 849 |
| 5 | CNRS | FR | 298 080 | 0 | 3 120 | 301 200 | 0 | 226 680 |
| 6 | KYOS | CH | 170 880 | 0 | 3 840 | 174 720 | 0 | 132 000 |
| 7 | Fraunhofer | DE | 983 778 | | 7 095 | 990 873 | 0 | 744 928 |
| 8 | CWI | NL | 192 429 | | 2 595 | 195 024 | 0 | 146 916 |
| 9 | Vocapia Research | FR | 338 880 | | 5 216 | 344 096 | 0 | 259 376 |
| 10 | Nielsen | DE | 430 625 | | 43 095 | 473 720 | 0 | 258 407 |
| | Total | | 3 495 300 | 0 | 484 523 | 3 979 823 | 0 | 2 998 338 |

**Figure 19 - SEMACORE budget**

### B 2.4.2 Details of other direct costs

*Travel expenses*

In order to contribute to overall spare use of energy and reduction of greenhouse gas emission, travel will be limited and replaced, as much as possible, by e-Meetings (audio/video conferences). A fixed travel expenses budget of 10.800 € is considered for each partner organisation, on the basis of four annual trips to consortium meetings averaging 900 € each (which could potentially accommodate 2 people travelling rather than 1).

Basic travel costs include participation to 4 consortium meetings (GA and PEB) per year plus additional WP-specific meetings, workshops, conferences especially in the context of dissemination, exploitation, collaboration with other projects.

An extra travel budget line of up to 10 K€ has been set aside to cover the travel expenses of the members of the External Advisory Board (see calculation details below).

The total travel expenses budget for the whole project is 114 400 €.

*Equipment expenses*

A budget line of 40 K€ has been established to cover the purchase of the computing and storage facilities that will support the implementation of the SEMACORE project plan. These hardware resources will be held and maintained at FRAUNHOFER in Bonn.

*Advisory board expenses*

Advisory Board members will not be financially compensated for their contribution to the project. However, in order to ensure their involvement, their travel expenses will be reimbursed.

With 2 meetings organised over the whole duration of the project (end of period 1 and end of period 2) and a unit travel cost of 1000 € for 4 EU advisors and 2000 € for 2 non-EU advisors, this results into a maximum amount of 16.000 € for EAB travel expenses. Anticipating a participation ratio of two thirds of the advisors at each meeting, this amount has been reduced to 10 000 € and has been integrated to the ERCIM budget.

*Dissemination expenses*

A lump sum of 10 K€ has been reserved on the ERCIM budget to cover dissemination expenses such as:

- participation to events and conferences,
- production of posters, flyers, brochures, printed reports,
- specific contribution to renowned publications,
- minor costs related to the web site,
- etc.

### B 2.4.3 Hardware resources committed by the consortium

Fraunhofer is committing components of its high-performance computing cluster infrastructure. The research cluster is part of a Big Data initiative and 'Living Lab' environment for real-time Big Data Analytics. It contains 35 nodes, each with a pair of up to 2.8 GHz, eight-core CPU and 1 TB of memory. For higher I/O-performance 3 nodes use a dedicated PCIe-based SSD cache optimization solution. Also the cluster contains a number of Nvidia Tesla K20M single graphic computing units delivering up to 9.2 teraflops of single precision floating point performance. Storage capacity of the cluster exceeds a total

of 100 TB. All nodes communicate over an InfiniBand network, but the management network uses a GigE LAN. The cluster is part of Fraunhofer's awarded Green IT computer center.

### B 2.4.4 Sub-contracts

SEMACORE will not use any sub-contractor for any part of its work plan.

Audit Certificates

The average cost for producing Certificates on Financial Statements (CFS) has been evaluated at 2k€ per unit for partners that will require such certificates. On this basis, a budget line of 8K€ has been requested to cover the anticipated audit costs. The expenses for CFS are declared as subcontracting costs in the management category.

# B 3. Impact

How much impact would have a tool available for policy makers or online service marketers capable to answer such type of questions, tapping directly on the online sources, with no manual research?

- *What fraction of videos in YouTube people in France and Germany talk about xenophobia in September 2013?*
- *Which Internet tablets were mentioned in blogs in UK and Spain during 2012 and what were the opinions on them.*
- *What are the educational contents available in Catalan?*

SEMACORE's goal is to provide ***an intelligent, integrated and comprehensive framework characterizing the web by cross-language multimedia knowledge mining capable to provide dynamic answers to those and other similar questions.*** SEMACORE will develop a workflow to implement text categorization, opinion mining and speech analysis for a new language with little effort.

## B 3.1 Strategic impact

SEMACORE responds to a fast growing interest from a variety of stakeholders, from the digital economy (online content, applications and service providers) to the information society realm (national, regional and international public sector). The solution to the challenges addressed by the project shall go beyond the immediate scope of the project, and lead to a new breed of products and insights. It will cover the measurement of speech and audio contents, the analysis of online behavior of users and the extensive and integrated use of analytics to capture knowledge and opinions. As such, the SEMACORE project will contribute to the following impacts from the ICT 2013 work programme.

| Expected Impact in WP | SEMACORE Contributions |
|---|---|
| Reinforced ability to extract, interpret and exploit information from **unstructured multilingual and/or multimedia sources** yielding actionable knowledge. | SEMACORE will provide language, content and opinion analysis for nine languages (Arabic, Catalan, Dutch, English, French, German, Italian, Spanish, Turkish). The approach ensures that the semantics of annotations are identical over the different languages and yields better results than the investigation of single languages. Text and transcribed speech from audio and video will be interpreted with the same semantic categories to exploit the generated information in three representative use cases. |
| Provide effective solutions that support multilingual business and interpersonal communication, and enable people to **access digital services in Europe's many languages**. | By analyzing the statements and opinions of users on specific issues in web pages, forums and blogs, SEMACORE will collect valuable information on products and services in a uniform format across languages. This information can be utilized by public agencies, private companies and can be provided to the public as a service in different languages which allows informed choices of citizens on digital services and public issues. In addition the investigations on language use in Europe give significant information on the resources accessible with languages with a smaller number of speakers allowing important decisions for supporting those languages. |
| Increased ability to operate new analysis algorithms to analyse, interact and visualize **extremely large volumes** of data in real time. | SEMACORE will develop an adaptive, scalable and dependable, real-time infrastructure for multimodal and multilingual content and opinion mining. It yields a real-time distributed programming framework that offers MapReduce functionality, provides end-to-end near real-time and reliable delivery of continuous streams of data. It combines a streaming engine for fast online analysis with a big data warehouse for analyzing and processing past data and annotations. This framework can be easily be adapted to new application contexts and scales effectively to new use cases. |

| | |
|---|---|
| Increasing ability to detect and exploit otherwise **hidden meaning across a range of applications**. | In contrast to search engines, SEMACORE will extract semantic information on the content categories, the meaning of keywords and their attributes. The analysis of the use and content of different languages gives important insights for an informed language policy. The investigation of opinions on enterprises and their products provides companies, citizens and customers with cross-media, cross-language and cross-national information for economic or policy decision making. |
| Strong participation of the **private sector** players, including SMEs, well above FP7 ICT average. | SEMACORE is committed to the private sector with Nielsen as a world leading market and media research enterprise as the main use case partner. Nielsen ensures the transformation of SEMACORE results to products focused to European and world markets. Four partners are SMEs which yield the access to business innovation in a fast growing market of contents, particularly in non text media and user generated content. |
| Increased capabilities to cope with multimedia content and everyday language with potential to **support real-life processes**. | SEMACORE will consider user's everyday language as found in emails, blogs, twitter, as well as in videos. It will analyze statements and opinions of users and aggregate for public and private stakeholders. The attitude and feelings of citizens with respect to specific topics can be elicited for decision making support. |

In addition to the expected impacts required by the 2013 Work Programme, SEMACORE will match strategic visions of EU towards Horizon2020, as mentioned in WP2013:

| **Strategic items in WP** | **SEMACORE Contributions** |
|---|---|
| Transforming our society through ICT developments in providing responses to **major societal challenges** | Connecting the next billion users and the wider diffusion of social networking and user generated contents, on the top of broadband infrastructures, will push up multilingualism in the democratization of the access to the Internet inside the global and European agenda. SEMACORE will contribute to European Union taking the lead in that new stage. |
| Preparing the expected launch of **Horizon2020**. | Call 10 is a transition call paving the ground for Horizon2020. SEMACORE is itself a transition project preparing for the next coming technological challenges of 2020 related to big data, analytics, semantic extraction and languages coverage with an inter-disciplinary approach favoring cross-fertilisation between technologies and applications and gathering different research and users constituencies. |
| Allowing people to access and use online content and services **across language barriers** | If social decision making is based upon pages accessible only from Search Engines indexes, then more than 90% of the web remains dark. SEMACORE will identify which language groups, national sites, and user groups are under-represented in search engines indexes, thus providing the means for tapping into these unexploited data sources. . This will constitute an historical breakthrough at the time when the lack of knowledge of Web contents allows myths and misconceptions to prevail[50]. There is a need to reveal the actual characterization of the web, language wise, so to untap social intelligence for multilingual online sources. |

### B 3.1.1 Other impact factors

#### *Multilingualism impact*

The impact of SEMACORE reaches out to Europe's citizens, businesses, industry and governments which are all, in some way, concerned by an Internet reflecting fairly the distribution of languages in the real world. SEMACORE addresses and provides responses to a major societal challenge: the emergence of multilingualism as a key factor of the future of the digital world and its implications on social behaviours, on democratic processes and on creativity. The building of solid linguistic indicators in the digital world

[50] How many Search Engine users and even professionals are duly aware of the fact that the index coverage has dropped from 80% which was the norm until 2006 into less than 10%? Are the ICT professionals clearly aware of the consequence of that situation? Are language policy makers aware of a situation and its consequences in terms of the future of languages in the online realm?

will definitively influence on policies and represents an important drive for development for the decades to come. Multilingualism is a fundamental element of knowledge societies. SEMACORE represents an important step in addressing the World Summit on the Information Society (WSIS)'s call to "*encourage the development of content and put in place technical conditions in order to facilitate the presence and use of all world languages on the Internet*"[51].

### *Market impact*

Much required marketing information (e.g., knowing user opinion) is very hard to capture, requiring expensive personal interaction such as calling users, or distributing and collating questionnaires. Thanks to SEMACORE providing tools for measuring user opinion on the web (especially in social networks), and from polling via the voluntarily installed SEMACORE plug-in, marketing agents will be able to ask specific questions to a wide audience across languages, as well as measure the dynamics of user opinions in the web. SEMACORE will enhance forecast capacity for content industry in languages hitherto ignored. SEMACORE will also provide the European market research industry with dynamic polling tools that will allow competing against established US competition: Google Trends, Zeitgeist. SEMACORE provides the hard facts basis for the strategic decision of the European in search engines market bringing unique data of the covering of actual indexes in terms of European languages. The language dimension emphasized by SEMACORE will give EU a strong advantage on important non-English emerging markets (China, Russia, Brazil, Vietnam and India). The extension of web-based Big Data mining to audio data will open new markets. As an example, SEMACORE will in the future permit a company to analyse the opinions towards the products of competitors and to decide whether or not to market its product in a country.
Finally social networks have changed the way information is delivered to the customers, shifting from traditional one-to-many to one-to-one communication. Opinion mining and sentiment analysis provided in SEMACORE offer the possibility to understand the user-generated comments, explain how a certain product or a brand is perceived and open an avenue to direct user centred marketing of products and services uniform across languages and nations.

### *Social impact*

SEMACORE provides a decisive step to innovate policy-making and eParticipation. In its use case on the multilingual analysis of web content and opinions for an international public agency it performs an open and transparent collection of the opinions of citizens in different countries on specific issues uniformly across languages. In the context of the EU this may be used for a democratic feedback on the policy cycle during the phase of policy design as well as policy monitoring and evaluation [Osimo 2012]. SEMACORE will open the door to new and user-centric measurement for producing indicators of the Information Society. Since current measurement is largely limited to the quantity of hardware and subscriptions, there is a clear lack of information on the content side, as well as on usage and applications. Eurostat has, with one of the partners of the consortium (DIALOGIC), started to conduct some preliminary research [Brennenraedts 2008] to produce indicators to analyze the digital behaviour of users. In a more extensive way, SEMACORE will contribute to the methods in which current indicators for the information society are produced working in close cooperation with the *Partnership on Measuring ICT for Development's* Task Group on Measuring the WSIS Targets[52]. Finally, SEMACORE will allow international organizations to improve governance effectiveness by a more detailed and timely knowledge about its constituencies.

[51] See Target 9 in:
http://itu.int/itunews/manager/display.asp?lang=en&year=2005&issue=09&ipage=wsis_targets&ext=html
[52] http://www.itu.int/ITU-D/ict/partnership/ composed by EUROSTAT, OECD, UNESCO/UIS, UNCTAD, UNDESA, UNECA, UNECLAX, UNESCAP and UNESCWA.

### B 3.1.2 European added value

SEMACORE's identification of under-indexed and dark language areas in the European community will feed the decision process for region-specific search engines, as well as alerting European language policy officials as to where efforts need to be taken. The European Community needs tools to deal with the costly language challenge of multilingualism; SEMACORE will provide inputs for a better management. In addition, SEMACORE will offer a tool for a better capture of Europe citizen's feedback on its programs and actions within predetermined axis of interpretation, as well as facilitating their graphic expression.

#### B 3.1.2.1 Interaction with other national and international research activities

The following is a list of EU and Nationally-funded relevant projects SEMACORE partners have participated or are currently participating in and their relation with SEMACORE:

| Partner | Project Name | Description | Comment |
|---|---|---|---|
| **FhG (coor.)** | **LIFT** – Local Inference in Massively Distributed Systems. FET-OPEN | Uses sketches and geometric monitoring for efficient mining of data streams | Results will be used in SEMACORE. No large-scale application, focus on basic research |
| **FhG** | **DataSim** – The Science of Data in the Area of Electric Vehicles. FET-OPEN | Big Data for improved traffic modelling and simulation, exploration of mobile phone data | SEMACORE will be informed of results. No language and real-time component; focus on basic research. |
| **FhG** | **DICODE**- Mastering Data Intensive Collaboration and Decision Making FP7-ICT2009-Call5 | Collaboration and decision making in data intensive settings; uses Hadoop, MapReduce, Mahout | No real-time streams, no multilingualism and no speech data. SEMACORE will be informed of results |
| **FhG** | **INSIGHT** - INtelligent Synthesis and Real-tIme Response using Massive StreaminG of HeTerogeneous Data | Manage emergency situations by analyzing information from sensors, Twitter and mobile phone connections. | No multilingualism and no multimodality. SEMACORE will be informed about the results. |
| **FhG** | **THESEUS** national project | Development of semantic analysis methods for text and multimedia documents. Results applied in the National Library and an industry partner. | Named entity recognition and relation extraction; no opinion mining, no multilingualism. Only small scale applications and search engines. |
| **UPC** | **FAUST** Feedback Analysis for User Adaptive Statistical Translation | UPC develops incremental learning techniques to improve MT performance using user feedback on system's output. | The developed adaptation methods will be used in SEMACORE. |
| **UPC** | **MOLTO** Multilingual On-Line Translation: Hibrid Approaches for Machine Translation | UPC develops hibridization techniques to combine the precision of rule-based systems with the coverage and robustness of statistical MT systems. | The acquired expertise will be useful for SEMACORE. |
| **UPC** | **XLIKE** Cross-lingual Knowledge Extraction. | TALP develops tools for multilingual semantic analysis of web documents, including non-standard language. | These tools will be used in SEMACORE to enrich documents with semantic information that will improve classification and sentiment analysis. |
| **LIMSI & VOC.** | **QUAERO** | Promote research and industrial innovation on technologies for automatic analysis & classification of multimedia & multilingual documents. | Results will be used in SEMACORE. The focus is on developing STT systems for all main European languages plus speaker identification. |
| **LIMSI** | **MEDIABOX** | Ameliorate results on STT for | Results will be used in SEMACORE. |

| | | | |
|---|---|---|---|
| **& VOC.** | Media watch | French and work on speeding up the recognizers. | |
| **LIMSI & VOC.** | **DARPA GALE** Global Autonomous Language Exploitation | Develop software technologies to facilitate the analysis interpretation of huge volumes of speech and text in multiple languages, Automatic processing engines distilled data, delivering pertinent, consolidated information to analysts. | LIMSI and Vocapia were partners on the BBN-led Agile team, developing speech recognition systems for the Arabic and Mandarin languages. |
| **LIMSI & VOC.** | **EDYLEX -** Enriching DYnamically LEXical resources in multilingual and multimodal apps | Study the dynamic acquisition of new lexical entries in existing lexicons used within linguistic processing systems | Results could be used in SEMACORE. Work on automatic updating of French and English vocabularies. |
| **LIMSI & VOC.** | **BABEL** - IARPA program | Agile & robust speech recognition rapidly applied to any language. | Provide effective search capability for analysts to efficiently process massive amounts of real-world recorded speech. |
| **KYOS** | **DEMONS -** Decentralised, cooperative and privacy preserving Monitoring for trustworthiness | Secure digital society with big data scenario can only function when trust in guaranteed in networks, services and applications. | Some results will be used in SEMACORE for security and privacy issues as well as regulatory analysis |
| **KYOS** | **EUINCOOP** Euro-India cooperation in computing systems | CSA project | Though not directly relevant, has some international aspects of interest |
| **KYOS** | **SEINIT** Security Initiative with IPv6 | IPv6 being a next generation IP, Security issues have to be considered. | This was an FP6 project, in which KYOS participated in intrusion detection solution development. |
| **MAAYA** | Study of French presence in the Internet | A study for OIF, follow-up of studies conducted since 1988. | Some methods and data may be useful for SEMACORE |
| **MAAYA** | Study of languages of France in the Internet | A national study for DGLF | Some methods and data may be useful for SEMACORE |
| **MAAYA member** | Observatory of languages & Cultures in the Internet | The observatory exists since 1988. | Some methods and data may be useful for SEMACORE |
| **MAAYA member** | Language Observatory Project | Measurement of languages in the Internet in Asia, Africa and currently Latin America | Some methods and data may be useful for SEMACORE |

The following table describes **recent EU projects on relevant topics**. These EU projects define the current state of the art. In the table we describe how SEMACORE differs from each and extends.

| **Name and Description** | **Relation to SEMACORE** |
|---|---|
| **COCKPIT -** Citizens Collaboration & Co-Creation in Public Service Delivery. COCKPIT aims to improve public service delivery decision making process by combining opinion mining, Service Science Management, and engagement of citizens. | COCKPIT will use opinion mining techniques over citizens' public service related interactions in blogs, forums, wikis, etc. No cross-lingual or cross-media analysis is foreseen, only 3 specific demonstrators at Greece, Italy and the Netherlands, respectively. |
| **NOMAD** - Policy Formulation and Validation through non-moderated crowd-sourcing. Nomad's vision is to provide decision-makers with fully automated solutions for content search, acquisition, categorisation/visualisation that work in a collaborative form in the policy-making arena. | NOMAD proposes multilingual opinion mining to extract political opinions. No text to speech transcription is used. In contrast to relying on "semantic classifiers" as in NOMAD the SEMACORE project will exploit the sequential and structural information of language by deep learning approaches offering a higher potential reliability. |
| **CROSSOVER** – Bridging Communities for Policy-Making. Crossover is a EU project building a | CROSSOVER defines a research roadmap for next generation policy-making and next generation policy---making |

| | |
|---|---|
| research roadmap for Policy-Making 2.0. | solutions. They discuss concepts handled in SEMACORE, for instance opinion mining and develop a roadmap, but do not concentrate on actively generating analysis methods. SEMACORE can utilize the recommendations of CROSSOVER. |
| **RENDER** - Reflecting Knowledge Diversity. RENDER will provide a comprehensive conceptual framework and technological infrastructure for enabling, supporting, managing and exploiting information diversity in Web-based environments. | RENDER aims to spot and assess the intensity of opinions expressed; and track topics along multiple sources, across data modalities and languages. RENDER does not consider a panel of volunteers to monitor language and content of information produced by users. In addition it does not measure language use in different countries and contexts. |
| **KHRESMOI** - Knowledge Helper for Medical and Other Information Users. Develop a multi-lingual multi-modal search and access system for biomedical documents. | KHRESMOI performs multilingual information extraction from biomedical documents but does not use opinion mining. |

#### B 3.1.2.2 External factors for impact achievement

None

#### B 3.1.2.3 Standards

To the extent possible the consortium will support EU policies to follow existing standards in international projects and to contribute to new standards in areas of human language technologies. For the representation of documents and the extracted annotations, a standard format will be used, e.g. CAS available from Apache UIMA.

## B 3.2 Plan for the use and dissemination of foreground

### B 3.2.1 Dissemination

The SEMACORE project has started from the requirements of indicators from a group of users composed of international and civil society organisations which have sought the support of a distinguished set of multi-disciplinary researchers capable of overcoming the actual limits of the state of the art to understand the structure of the web contents. It is then logical that the exploitation and dissemination of the results of the research occupies an important part of the project.

**For Scientific partners**
The project will naturally contribute to the advancement of science by publishing. Project members are expected to produce publications at major scientific journals, and present their work at international conferences and workshops. The project will also organize special sessions at major conferences, and satellite workshops collocated at major conferences, for maximum visibility. Beyond the common scientific process, the consortium will participate in international benchmarking exercises and campaigns to compare the advances and innovations of SEMACORE to the state of the art.

**For the consortium**
It is considered essential to communicate SEMACORE results to the public as well. This activity will involve communicating the goals and activities of the project in laymen's terms so to reach out to the public and to public bodies via public demos, Press Conferences and participation in Trade-Shows and Fairs. Furthermore, the marked interest of International Organisations such as UNESCO, ITU or OIF and

the linkage with international cooperation frameworks such as WSIS and IGF, will deliberately orient the exploitation of the results of the project towards public policies in the field of Information Society.

**International events**
Two workshops will be organised to present the outcomes of the project and share experience with other experts working in those fields.

### B 3.2.2 Exploitation Plans

There are two strategic lines for exploitations of the results of SEMACORE, one oriented towards linguistic policies and of interest to national or international organizations; the second one oriented towards online economy and business driven by the market forces.

**Policy oriented exploitation plans**

Language policies cover a fairly heterogeneous set of activities ranging from promotion, protection to revitalization. The **indicators** are the fundamental guiding tool for language policies. Although many national or international organisations offer statistics, surveys, studies, and other types of indicators, they differ from one institute to other providing glaring inconsistencies. The observation of the evolution of languages in cyberspace is no exception to this rule. SEMACORE, while contributing to the creation of indicators on languages in cyberspace, will also put in parallel the existing information about weights on the real place of language in society, opening new exploitation possibilities for creating universal indicators for language planning criteria and decision making. The stakeholders which represent the users from the institutional side and whose interests are represented by MAAYA are already planning an exploitation project outside the FP7 framework which will build upon the outcomes of SEMACORE and systematize the produced indicators. SEMACORE will organize a set of preparatory activities to facilitate the triggering of this posterior stage; Task 8.2 is specifically dealing with those exploitation activities which encompass activities to ensure the sustainability of the production of indicators and a roadmap for further researches in that field.
It is increasingly recognized that the emergence of new and complex problems, e.g. the financial crisis, requires government to increasingly collaborate with non-governmental actors in the understanding and in the addressing of challenges [Osimo 2012]. SEMACORE will exploit its approach on the cross-language collection of citizen's opinions on public and societal issues and offer custom-made solutions to international public agencies, administrations, and NGOs. This approach can take into account the mutual influences between people in social networks and elicit the effect of regional, cultural, national and social backgrounds. It allows a very fast, uniform and comprehensive survey of opinions, which give valuable input to the stakeholders to improve policy formulation and public relations.

**Business oriented exploitation plans**

Beyond the exploitation plans outlined in T8.4 and beyond the set of use cases designed in WP7 to demonstrate concrete exploitations of the researches outcomes, SEMACORE opens a wide range of exploitation opportunities in the public and private sector which are described below.

#### B 3.2.2.1 Target Organizations of the SEMACORE Exploitation

***Political Administrations and Organizations:***
A number of recent political situations show the value and effectiveness of Web 2.0, in relation between citizen and public administrations. This is the case of different electoral processes (i.e. the Presidential Election in the United States of America, the current primaries in the US, as well as some elections in European countries); also in legislative processes (the increasing use of tweets and Facebook encouraging to lobby for or against a specific legislation being under discussion), or in spontaneous

citizen movements such as it happened in the explosion of Arab spring or in the various occurrences of 'indignados'. The web is a tool for free thought and for free speech and Web 2.0 has accelerated and democratized the ownership of information by the citizens. Being able to tap into this massive knowledge and opinion base would allow administrations and political parties to better address the needs and wants of their constituents with fewer resources. For European governments and institutions it is especially valuable to elicit the opinion of European citizens in a consistent way. The push in administrations for Open Data will also open windows of creative opportunities for the exploitation of the SEMACORE platform.

***Prevention and Health:***

As widely admitted, the prevention of disasters, disease, and crime are topics which are globally relevant. Many health care providers, such as national health systems, hospitals and assisted care facilities are searching for ways to reduce spiralling costs of care. National systems of security, insurance companies, justice and other state bodies or private entities for prevention need rapid interaction with the citizen. An advanced feedback evaluation based on the opinions of citizens and professionals and performed with the techniques of the SEMACORE project would assist in reducing costs and improving standards of care and prevention and may reduce misconceptions on health and diseases.

***Publishing/Media Companies:***

Worldwide, the publishing industry is one of the industries which are suffering the worst under the current economic crisis. Some publishing houses have integrated Web 2.0 features such as blogs, forums and twitter into their existing legacy systems. This allows for customer/reader feedback. This creates a problem of measuring and analyzing the feedback, which is not the core business for the publishing companies. The platform developed in the scope of the SEMACORE Project would give publishing companies the basis technology to easily measure and analyze feedback from their customers on the basis of blog comments, forum entries and/or tweets (Twitter) in a uniform way in different languages. This would then give publishers the ability to tailor their content offering to what their customers/readers find interesting, identify experts from the community and generally engage their customer base in new and interesting ways.

***Market Research Firms:***

Traditionally, market research firms have relied on a variety of techniques to record, aggregate and summarize qualitative and quantitative data. The problem for international markets is the fragmentation of the research by language barriers. Within the scope of the SEMACORE project, cross-lingual techniques will be developed and applied to collected web contents. The technology will yield comparable results for the different languages (both text and audio) thus providing a unified view on the perception of a product or brand. The application is analyse customer experience with specific products or services, and to measure customer satisfaction. The approach potentially generates more statistically representative results but can be combined with traditional surveys. In addition the system enables Voice of Customer and Customer Experience Management professionals to automatically extract specific attributes of customer experience that impact Customer Satisfaction scores. The approach is able to generate Competitive Marketing Intelligence by producing scores for competing products/services.

Another problem which the project can solve for market research companies is that of data volume. Such companies can already sift through the mountain of data available on the web, but at a large cost due to the fact that it is mostly entered manually. Thus, the SEMACORE project allows for quicker and more accurate collection, preparation and analysis of information for market researchers. Finally the detailed knowledge of customer opinions provides the opportunity to initiate customer-specific marketing of products and services, where customers receive tailor-made offers which correspond to their specific needs. As Nielsen, the globally largest market research enterprise is a partner of the project, the SEMACORE approach will especially be geared to up to date and relevant solutions.

***Advertising and communication companies:***

The advertising and communication sector has been hit especially hard during the past economic downturn. The SEMACORE project will enable companies and their customers to benefit from multilingual analysis modules, allowing advertising and communication companies to see what is being said online about their customers and then customize ad placement, marketing campaigns, brand development and slogans. Furthermore, companies can be alerted to problems such as online brand bashing, where unfounded rumors about their products or services are posted online. Yet another facet is product development, where communication companies can work with their customers to help develop products or features of existing products based on input obtained from the SEMACORE platform.

***Search Engine Providers:***

SEMACORE will produce a new measure of the content and makeup of the Internet. One, large untapped source of information on the Internet concerns the currently ignored content of audio-visual streams. SEMACORE will allow the creation of a new generation of search engines including the language factor in audiovisual searches. The new geography of the Internet that will be brought to light will also illustrate where search engines can profitably devote energies to index unexplored content brought to light by SEMACORE from the huge dark web. Finally, SEMACORE will open the possibility of a new type of search engine products oriented towards opinions instead of plain text.

#### B 3.2.2.2 SEMACORE Selling Points

The SEMACORE solution has the following **unique advantages**:

- The analysis of contents and opinions is done for different languages in a uniform way yielding comparable results covering many countries.
- Text, audio and video sources are covered and analysed in a uniform manner. Therefore the results for each modality can be combined in a consistent way and provide a unified picture of web content. For instance, we are not aware of opinion mining and content analysis for many languages integrating the results from text and speech analysis.
- SEMACORE provides an integrated coverage of complementary personal restricted web content (email, social networks) by an integrated panel of individual volunteers.
- Scientific and commercial partners of the SEMACORE project, especially Nielsen - the globally largest market research firm - have an outstanding reputation. The SEMACORE consortium can financially profit through different **revenue streams** (custom analyses for business partners, extension of the consortium members' own line of business, Data as a Service, multilingual data sources enriched with meta-data, consultancy services, Application Service Providing).
- Finally, the SEMACORE platform will be a real-time distributed programming framework that offers MapReduce functionality and provides end-to-end near real-time and reliable delivery of continuous streams of data. It will serve as a basis for other companies to develop additional technological and sector specific modules, creating a rich ecosystem of business solutions increasing the overall value of the platform.

#### B 3.2.2.3 Particular exploitation targets and plans of the project partners

The partner consortium has strong ties with contacts in large, international media conglomerations, leading business enterprises as well as in the public sector in areas such as healthcare, government and academia. The SEMACORE platform will provide services to a wide range of public and private entities, as well as consumers. Project partners include market leading organisations in their respective fields. This combination of academic, public and industrial partners is perfectly suited for working with existing customers and collaborators in a variety of industry and public segments to develop the SEMACORE platform for market use

***Fraunhofer***

FHG is carrying out contractual research work for many major companies in Germany. As an applied research organization, the core mission of FHG is to offer the SEMACORE technology to these companies, thus promoting the uptake of intelligent information management technology in commercial settings. FHG is planning to exploit the content mining technology developed in this project in its commercial projects, in particular in fields that face the problems of large and distributed amounts of data. In particular, FHG will investigate the use of SEMACORE's results in existing collaborations in the business areas of online advertisement, Smart Semantics, media company support and fraud detection.

***MAAYA***

MAAYA is preparing, with OIF support, an ambitious follow-up exploitation of the SEMACORE project in cooperation with UNESCO, ITU. It will take over, extend and systematize a selection of the indicators produced by the project, organize the dissemination of those results and exploit them in a non-profit perspective in different manners from workshops organized in some countries to a compound of strategy papers for language development and protection in cyberspace and for the support of multilingualism in the Digital World. There is absolutely no dependency from the SEMACORE project to this subsequent project, only the firm motivation of a set of institutional partners to move ahead a common agenda for language indicators construction in cyberspace. Negotiations will be conducted during T8.2 with some of the partners such as Fraunhofer, Dialogic or Vocapia to cooperate for the extension of data production after the end of the SEMACORE project.

***Vocapia***

Vocapia Research develops leading edge speech processing technologies. SEMACORE will allow Vocapia to extend their offering of languages, so as to cover a larger potential client base. Vocapia expects to improve their language identification technology in particular with the identification of languages in multilingual documents. Vocapia will also have the opportunity to develop products for new markets based on automatic opinion mining and sentiment analysis applied to multimedia and multilingual documents.

***Nielsen***

SEMACORE enables the industry to understand the consumer so that the industry is enabled to identify trends, demands and wishes of consumers and to provide the related products and services. The digital media is especially supporting small and medium enterprises to be innovative and creative and SEMACORE will provide them with relevant information for their business. Nielsen plans to exploit the SEMACORE system and its analysis tools for trend analysis, identification of new market places, new product and service development based on consumer requirements, Identification of purchase drivers and market sizes, analyzing key factors for improvement of product and service quality, tracking systems for products and services based on consumer feedback, getting insights into sentiment of product and services and tracking of issues and challenges.

### B 3.2.2.4 Overall Marketing Strategy and Perspectives

To ensure the sustainability of the SEMACORE project, consortium partners will formulate a detailed overall exploitation plan based on market analysis. As presented above, the industry partners in the SEMACORE project already operate successfully in their respective markets and sectors. This deep industry understanding will be drawn upon to assist in marketing the results and achievements of the SEMACORE project. The exploration and dissemination of the SEMACORE project will be based on modern marketing and communication best practices, using both online tools (portal, web 2.0, demonstrators) and offline supports (flyers, white papers, scientific Journals, presentations, conferences, workshops, trade fairs …). Further marketing and business development activities will be geared toward the identification of strategic partners in the subsequent geographic areas.

### B 3.2.3 Management of intellectual property

Proper intellectual property protection will be considered, also under the perspective of possible copyright protection, patentability and any other king of intellectual property protection, in relation to software and more generally any kind of know-how that will be produced in the course of the project, as a result of the same or that will be comprised in the outputs of the project. To this regard, know-how and any kind of intellectual property right developed in relation to a specific stage of the project or as an output of the same will be protected through appropriate procedures and agreements established among the members of the Consortium. Management and protection of knowledge and intellectual property will be eased, within the project, by the tight interaction with KYOS. From the early stage until the end of the project time, the legal department of the partners will provide continuous assistance as to management of the knowledge produced and protection of intellectual property rights in any way arising or connected with the project. The partners will determine the appropriate knowledge management procedures and rules within the Consortium at the various stages of work and thereafter, especially for what concerns the innovation aspects of this project. The Consortium Agreement will devote specific and significant attention to the issue of intellectual property rights management. In the Consortium Agreement they will be defined and specified procedures and rules for a proper handling, ownership, managing, protection and granting of the knowledge and of any relevant intellectual property rights, in any way produced and of any kind, with regards to both internal usage for scopes within the project frame, usage outside the project during the project time frame and usage after project completion.

# B 4. Ethical Issues

Gender issues

The SEMACORE consortium is well aware of the recent efforts in Europe aimed at fostering equality for women and for ensuring equal opportunities for persons with disabilities[53] and other minorities, including linguistic. Although all of the partners in the SEMACORE consortium are equal opportunity employers, we are aware as a community that women are generally underrepresented in the scientific and engineering fields upon which the SEMACORE technology will repose. Some of the research domains covered by SEMACORE already have a relatively equal gender balance - such as linguistics and human factors, while in computer science and engineering the percentage of women is substantially lower. The project will encourage the involvement of more women experts in the project development.

One way in which SEMACORE can help remediate the situation is by having the senior women and minorities serve as role models, thus encouraging junior researchers and students to pursue careers in scientific and engineering fields. At the partner sites women currently represent between 15 and 40 percent of the research staff. We will actively encourage female doctoral students to carry out research related to the SEMACORE project, and some sites will offer financial support for such studies. We will also encourage women to visit the other partner sites to forge deeper relationships with other women working on similar research problems.

The research environments at the partner sites are conducive to flex-time, so that researchers can balance work and family constraints. We will try to organize consortium meetings so as to satisfy both work and family situations. The dates and host location will be chosen to facilitate travel and to the extent possible minimize disturbance to family life (avoiding school vacation, weekend meetings or travelling on weekends, minimizing the nights away from home).

While the technology developed in SEMACORE will on average perform equally well for users of both genders, there will inevitably be individual differences. Testing will be carried out in situations representing both genders.

To the extent possible, SEMACORE will follow the recommendations of the European Technology Assessment Network (ETAN) report on promoting gender equality in scientific research. The report, entitled "Science policies in the European Union: Promoting excellence through mainstreaming gender equality"[54].In particular, the report provides recommendations for good practice in equal opportunity recruitment, and to promote gender equality in decision making processes and dissemination activities.
The project management will:

- Adopt the appropriate measures encouraging women participation in the management of the project, in order to achieve a balanced consortium
- Support the implementation of the recommendations produced by the European Technology Assessment Network (ETAN) on the development and production of statistics and indicators, about the situation of women in scientific research

Project-wide approach to ethical issues

[53] W3C web content accessibility guidelines (WCAG)
[54] http://www.cordis.lu/etan/home.html

Since the project is engaged in cross-domain analytics, ethical issues are to be considered from the beginning of the project with clearly defined guidelines. Some fundamental practices of concern include: the potential availability of users' data to third parties for commercial, surveillance or data mining purposes; the ability of third-party applications to collect and publish user data without their permission or awareness; the use of automatic 'opt-in' privacy controls; the capacity of facial-recognition software to automatically identify persons; the track of online user activities; the potential use of location-based for stalking or other illicit monitoring of users' physical movements; the sharing of user information or patterns of activity with government entities; and to encourage users to adopt voluntary but imprudent, ill-informed or unethical information sharing practices, either with respect to sharing their own personal data or sharing data related to other persons and entities.

The major ethical issue in SEMACORE is to inform users about legal implications and ethical measures taken to ensure the protection of personal data as the date collection will take place throughout different activities (WP's) in the project. The partners involved in SEMACORE will prepare guidelines on ethical use of information for the project, describing how data will be collected, why and how it will be used. The guidelines will include concrete recommendations for the exploitation and integration of the research outcomes at the national level. In addition, the end user will receive specific information (End User License Agreement/Terms of Use) regarding the ethical aspects of the project. Upon request and aiming to maintain the highest level of transparency, the users will receive the executive summary of the project outcomes.

A special committee (Virtual Secretariat) will be established for all questions related to the legal and ethical issues, particularly privacy (linked to Task 4.3). The policy dealing with privacy issues will be aligned with the European directive guidelines[55] and participants/users will be informed about the objectives of the project and personal information will be held confidentially.

The fundamental principles outlined in various EU and UN international normative documents as human dignity, integrity of the person, the right to privacy - just to mention the most prevalent ones - will be fully respected and promoted within SEMACORE. All legal and ethical requirements of the member states where the project is implemented will be fulfilled.

Specific action with regards to ethical issues

As part of WP4 'Societal Issues', SEMACORE addresses specifically all legal, ethical and regulatory matters (T4.2) and all privacy protection and security concerns (T4.3) associated with such a project.

SEMACORE does not raise any specific Ethical issue. Should any ethical matter arise during the project life time, the project management will promote the assessment by a recognised ethics committee in consultation with EC.
All participants in SEMACORE will conform to the legislation and regulation in force in their respective countries. The concerned rules to be observed are:

- The charter of Fundamental rights of the EU
- Council directive 95/46/EC of Oct. 1995 on protection of individuals with regard to the processing of personal data and on the free movement of such data.

As a fundamental complement of the technical requirement specification, a thorough analysis and profiling of the legal and regulatory framework will be provided, including both i) privacy and data security requirements set forth at a European level, and ii) privacy laws in selected EU countries. These will not be static profiles but will also take into account the position held by local data protection Commissioners and actual practice of the selected countries (in some territories, actual practice differs

[55] 2002/58/EC of the European Parliament and of the council concerning the processing personal data and the protection of Privacy in electronics communication sector; directive on privacy and electronic communications

from written law). A similar review of relevant law enforcement legislation will be undertaken to identify additional obligations and uniform technical models and standardised best practices. Rather than just provide an overview of all the rules in all the data protection laws and secondary rules and regulations, aim of this activity is to describe in a comparative and analytical way the laws in the selected Member States, in order to provide technical requirements which will ensure that the developed product will comply with the relevant rules of the EU regulatory regime. This will assist to get insights and solutions to achieve a high level of integration between technical concepts and European laws and regulatory provisions.

| **Ethical issues table** | YES | NO |
|---|---|---|
| **Informed Consent** | | |
| Does the proposal involve children? | | ✓ |
| Does the proposal involve patients or persons not able to give consent? | | ✓ |
| Does the proposal involve adult healthy volunteers? | | ✓ |
| Does the proposal involve Human Genetic Material? | | ✓ |
| Does the proposal involve Human biological samples? | | ✓ |
| Does the proposal involve Human data collection? | | ✓ |
| **Research on Human embryo/foetus** | | ✓ |
| Does the proposal involve Human Embryos? | | ✓ |
| Does the proposal involve Human Foetal Tissue / Cells? | | ✓ |
| Does the proposal involve Human Embryonic Stem Cells? | | ✓ |
| **Privacy** | | |
| Does the proposal involve processing of genetic information or personal data (eg. health, sexual lifestyle, ethnicity, political opinion, religious or philosophical conviction) | YES | |
| Does the proposal involve tracking the location or observation of people? | YES | |
| **Research on Animals** | | |
| Does the proposal involve research on animals? | | ✓ |
| Are those animals transgenic small laboratory animals? | | ✓ |
| Are those animals transgenic farm animals? | | ✓ |
| Are those animals cloned farm animals? | | ✓ |
| Are those animals non-human primates? | | ✓ |
| **Research Involving Developing Countries** | | |
| Use of local resources (genetic, animal, plant etc) | | ✓ |
| Impact on local community | | ✓ |
| **Dual Use** | | |
| Research having direct military application | | ✓ |
| Research having the potential for terrorist abuse | | ✓ |
| **ICT Implants** | | |
| Does the proposal involve clinical trials of ICT implants? | | ✓ |
| **I CONFIRM THAT NONE OF THE ABOVE ISSUES APPLY TO MY PROPOSAL** | **NO** | |

**Explanation**

The user-centric measurement of the project (WP6) potentially involves the processing of personal data and the tracking of the location of people. The first issue is the flip side of the unique ability of this type of measurement, namely to be able to measure social conduct in action (that is, language in real use). The second issue is inherent to the inclusion of mobile devices as they are the nexus of measurements.

We guarantee the privacy of our respondents by two complementary measures: (1) in the design of the measurement client and (2) validation by an independent third party.

Privacy issues may arise in the client for three different aspects: language determination, background information, and location information. We now discuss how we take privacy into account in the client's design, for each of these three dimensions.

In order to determine the language that is being used by the respondent, we send pieces of (potentially personal) information from the respondent's site to a central server. This involves sending a limited number of randomly taken nouns, adjectives or verbs over an encrypted connection. First, it is unlikely to derive from the centrally collected data any real personal data, since the data collection takes place at random. Second, once the algorithm at the server has identified the language the original input string is shredded.

During the first installation of the client respondents are asked to provide a limited number of general background characteristics. These background variables are the basis for analysis of the output data later on. To safeguard the privacy of the respondent, we use a randomly generated ID that is stored in a hash table. Given the sizeable number of IDs, even the use of rainbow tables will not enable reverse engineering of the IDs.

Finally, with regard to the location of the respondent, we will not collect or store the location information that is automatically being generated by the mobile device. In theory, though, should the background information store too fine grained geographical information, the combination of several background variables with data received could still identify a particular respondent. We counter this possibility by restricting geographic data to a very aggregate geographical location in the background variables (country), and broad categories that have at least several hundreds of respondents in one subclass (thus ensuring k-anonymity).

To ensure that all proper security measures are taken to safeguarded the privacy of our respondents, we have established a strict divide between the entity who will develop the user-centric client (Dialogic) and the entity (Kyos, a respected IT security firm) who will investigate the programming code, perform vulnerability tests, audit the security procedures that are taken at Dialogic and in the consortium as a whole, and provide advice on strong authentication and intrusion detection (at the server side).

In addition, WP 4 will develop the End User License Agreement/Terms of Use) for legal implications and inform the users about ethical measures to be taken to ensure the protection of personal data and the statistical usage of the collected data within the framework of the project. This work package will coordinate with all the partners guidelines on ethical use of information for the project describing how data will be collected, and why and how it will be used. The fundamental principles outlined in various EU and UN international normative documents as human dignity, integrity of the person, the right to privacy, etc. will be analyzed in order to be fully respected and promoted and feed task in the WP 4.

# B 5. Annexes

## B 5.1 References

Note: Authors who are **members of the consortium** partner organisations are shown in bold.

[Aggarwal 2012] Charu C. Aggarwal, Chen Xiang Zhai (eds.) (2012): Text Mining. Springer 2012.

[Ahmadi 2012] **B. Ahmadi, K. Kersting, S. Natarajan** (2012): Lifted Online Training of Relational Models with Stochastic Gradient Methods. Proc. ECML PKDD 2012. Selected as one of the best papers.

[Albuquerque 2010] Albuquerque A., Esperança, J.P. (2010): El valor económico del portugués: lengua de conocimiento con influencia global (ARI)

[Alper 2011] Basak Alper, Huahai Yang, Eben Haber, Eser Kandogan (2011): OpinionBlocks: Visualizing Consumer Reviews. IEEE Workshop on Interactive Visual Text Analytics for Decision Making at VisWeek 2011.

[Andrienko 2012] **Gennady Andrienko, Natalia Andrienko**, Harald Bosch, Thomas Ertl, Georg Fuchs, Piotr Jankowski, Dennis Thom (2012): Discovering Thematic Patterns in Geo-Referenced Tweets through Space-Time Visual Analytics. Submitted to IEEE Computing in Science and Engineering,

[Banea 2010] Carmen Banea, Rada Mihalcea, Janyce Wiebe (2010): Multilingual Subjectivity: Are More Languages Better? Proc. EMNLP 2010.

[Banea 2011] Carmen Banea, Rada Mihalcea, Janyce Wiebe (2011): Multilingual Sentiment and Subjectivity Analysis, in Multilingual Natural Language Processing, Editors Imed Zitouni and Dan Bikel, Prentice Hall, 2011.

[Benamara 2012] Farah Benamara, Baptiste Chardon, Yannick Mathieu, Vladimir Popescu, Nicholas Asher (2012): How do Negation and Modality Impact on Opinions? Proc. of the ACM Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics (ExProM 2012)

[BenZeghiba 2009] **Mohamed Faouzi BenZeghiba, Jean-Luc Gauvain** and **Lori Lamel** (2009): Language Score Calibration using Adapted Gaussian Back-end, Interspeech'09, pp. 2191-2194, Brighton.

[BenZeghiba 2012] **Mohamed Faouzi BenZeghiba, Jean-Luc Gauvain** and **Lori Lamel** (2012): Fusing Language Information from Diverse Data Sources for Phonotactic Language Recognition Odyssey 2012, Singapore, June 2012.

[Bi 2011] Wei Bi, James T. Kwok (2011): Multi-Label Classification on Tree- and DAG-Structured Hierarchies. ICML 2011.

[Bockermann 2012] C. Bockermann, Hendrik Blom (2012): The Streams Fraumework. Technical report University of Dormund.

[Boyd-Graber 2010] Jordan Boyd-Graber, Philip Resnik (2010): Holistic Sentiment Analysis Across Languages: Multilingual Supervised Latent Dirichlet Allocation. Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP 2010).

[Brennenraedts 2008] **Brennenraedts, R.,** C. Holland, R. Batenburg, P. den Hertog, **R.A. te Velde**, S. Jansen, S. Brinkkemper (2008) :Go with the dataflow! Analysing the Internet as a data source (IaD). Final Report prepared for the Dutch Ministry of Economic Affairs. Utrecht: Dialogic.

[Brennenraedts 2012] **Brennenraedts**, **R. R.A. te Velde** (2012): Internet as data source. Feasibility Study on Statistical Methods on Internet as a Source of Data Gathering. Final Report prepared for the European Commission, DG Communications Networks, Content & Technology. Utrecht: Dialogic.

[Brown 1990] Brown, P. F., Cocke, J., Pietra, S. A. D., Pietra, V. J. D., Jelinek, F., Lafferty, J. D., Mercer, R. L.,, Roossin, P. S (1990): A statistical approach to machine translation. Computational Linguistics 16(2), 79-85, 1990.

[Brown 1993] Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., Mercer, R. L. (1993): The mathematics of statistical machine translation: parameter estimation. Computational Linguistics 19(2), p. 263-311, 1993.

[Brummer 2012] Niko Brummer, Sandro Cumani, Ondrej Glembek, Martin Karafiat, Pavel Matejka, Jan Pesan, Oldrich Plchot, Mehdi Soufifar, Edward de Villiers, Jan Cernocky (2012): Description and analysis of the Brno276 system for LRE2011. Odyssey 2012, Singapore, June 2012.

[Burget 2006] L. Burget, P. Matejka, J. Cernocky (2006): Discriminative Training Techniques for Acoustic Language Identification, Proceedings of ICASSP'06, Vol. 1, pp. 197-200, Toulouse, 2006

[Carreras 2009] **Xavier Carreras**, Michael Collins (2009): Non-projective Parsing for Statistical Machine Translation. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP), 200-209, Singapore, 2009.

[Castaldo 2007] Fabio Castaldo, Daniele Colibro, Emanuele Dalmasso, Pietro Laface, Claudio Vair (2007): Compensation of Nuisance Factors for Speaker and Language Recognition, IEEE Trans. Audio, Speech and Language Processing, Vol 15(7), pp. 1969-1978, 2007.

[Chiang 2005] Chiang, D. (2005): A hierarchical phrase-based model for statistical machine translation. In: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005), 263-270. Ann Arbor, Michigan, USA, 2005.

[Choong 2009] Chew Y. Choong, Y. Mikami, C. A. Marasinghe, S. T. Nandasara (2009): Optimizing n-gram Order of an n-gram Based Language Identification Algorithm for 68 Written Languages, The International Journal on Advances in ICT for Emerging Regions 2009 02 (02) : 21 - 28

[Clements 2009] Maarten Clements, **Arjen P. De Vries**, and Marcel J. T. Reinders. 2010. The task-dependent effect of tags and ratings on social media access. *ACM Trans. Inf. Syst.* 28, 4, Article 21 (November 2010).

[Collobert 2012] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, Pavel Kuksa (2012): Natural Language Processing (Almost) from Scratch. Journal of Machine Learning Research Vol. 12 (2011)

[Davison 1997] Davison, A. C., Hinkley, D. V. (1997): Bootstrap methods and their application. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

[Dey 2008] Lipika Dey, S K Mirajul Haque (2008): Opinion mining from noisy text data. In Proceedings of the Second Workshop on Analytics for Noisy Unstructured Text Data (AND-2008). 2008.

[Dhar 2012] Vasant Dhar (2012): Data Science and Prediction (March 29, 2012). NYU Stern School of Business CeDER-12-01.

[Frank 2012] J.R. Frank, M. Kleiman-Weiner, D.A. Roberts, F. Niu, C. Zhang, C. Re, I. Soboroff. Building an Entity-Centric Stream Filtering Test Collection for TREC 2012. In Proceedings of the Twenty-First Text Retrieval Conference (TREC 2012). To appear.

[Graddol 2007] Graddol, D. (2007): English Next. Why Global English may Mean the End of "English as a Foreign Language". The British Council & The British Company (UK) Ltd. 2006 (new edition 2007).

[Harper 2008] Mary. P. Harper, Michael Maxwell (2008): Spoken Language Characterization, Springer Handbook of Speech Processing, Chap. 40, pp. 797-807, 2008.

[ITU 2009] ITU (2009): Measuring the Information Society: the ICT Development Index, ISBN 92-61-12831-9, 2009

[Jakob 2010] Niklas Jakob, Iryna Gurevych (2010): Extracting Opinion Targets in a Single- and Cross-Domain Setting with Conditional Random Fields. Proc. EMNLP 2010.

[Kenny 2007] P. Kenny, G. Boulianne, P. Ouellet, P. Dumouchel (2007): Joint Factor Analysis versus Eigenchannels for speaker recognition, IEEE Trans. Audio, Speech and Language Processing, 15(4), pp. 1435-1447, May 2007.

[Kersting 2012] **K. Kersting** (2012): Lifted Probabilistic Inference. Proc. ECAI-2012. Invited Talk at the Frontiers of AI Track

[Kim 2012] Sungchul Kim, Kristina Toutanova, Hwanjo Yu (2012): Multilingual Named Entity Recognition using Parallel Data and Metadata from Wikipedia. Proc. ACL 2012.

[Koehn 2003] Koehn,P., Och, F. J., Marcu, D. (2003): Statistical phrase-based translation. In Proc. of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL). Edmonton, Canada, 2003.

[Koehn 2007] Koehn, Philipp, Hoang, Hieu (2007): Factored Translation Models. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), 868-876, Prague, Czech Republic, 2007.

[Lamel 1995] **Lori Lamel** and **Jean-Luc Gauvain**, A Phone-based Approach to Non-Linguistic Speech Feature Identification, Computer Speech and Language, 9(1):87-103, January 1995.

[Lamel 2002] **Lori Lamel**, **Jean-Luc Gauvain**, **Gilles Adda** (2002): Unsupervised Acoustic Model Training. In Proceedings of ICASSP, pages 877-880, Orlando, May 2002.

[Lamel 2010] **Lori Lamel, Bianca Vieru** (2010): Development of a Speech-to-text transcription system for Finnish. In The second International Workshop on Spoken Languages Technologies for Under-resourced languages (SLTU10), pages 62-67, Penang, Malaysia, May 2010.

[Lefever 2010] Els Lefever, Veronique Hoste (2010): SemEval-2010 Task 3: Cross-Lingual Word Sense Disambiguation. Proc. Int. Workshop on Semantic Evaluation 2010.

[Lefever 2012] Els Lefever, Veronique Hoste, Martine De Cock (2012): Discovering Missing Wikipedia Inter-language Links by means of Cross-lingual Word Sense Disambiguation

[Liu 2008] Liu, D., Gildea, D. (2008): Improved Tree-to-String Transducer for Machine Translation. In: Proceedings of the Third Workshop on Statistical Machine Translation, 62-69. Columbus, Ohio, USA, 2008.

[Liu 2012] Bing Liu (2012): Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers.

[Liu 2012a] Gang Liu, Chi Zhang, John H L Hansen (2012): A Linguistic Data Acquisition Front-End for Language Recognition Evaluation Odyssey 2012, Singapore, June 2012.

[Malouf 2008] Robert Malouf, Tony Mullen (2008): Taking sides: User classification for informal online political discourse. Internet Research. 18:177-190.

[Marcu 2006] Marcu, D., Wang, W., Echihabi, A., Knight, K. (2006): SPMT: Statistical Machine Translation with Syntactified Target Language Phrases. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP), 44-52, Sydney, Australia, 2006.

[Marz 2011] Nathan Marz (2011): Twitter storm framework, 2011. URL: https://github.com/nathanmarz/storm/.

[Marz 2012] Nathan Marz, James Warren (2012): Big Data - Principles and best practices of scalable realtime data systems. Manning Publications.

[Maurais 2003] Maurais, J., Morris, M. A (2003): Languages in a globalising world. Cambridge: Cambridge University Press

[Mimno 2009] David Mimno, Hanna M. Wallach, Jason Naradowsky, David A. Smith, Andrew McCallum (2009): Polylingual Topic Models. Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP 2009).

[Mukherjee 2012] Arjun Mukherjee, Bing Liu (2012): Mining Contentions from Discussions and Debates. Proc. KDD'12, 2012.

[Nesreen 2012] Nesreen K. Ahmed, Jennifer Neville, Ramana Kompella: Network Sampling: From Static to Streaming Graphs, Purdue University, 2012

[Neumann 2011] **M. Neumann, B. Ahmadi, K. Kersting** (2011): Markov Logic Sets: Towards Lifted Information Retrieval Using PageRank and Label Propagation. Proc. AAAI 2011.

[Ni 2011] Xiaochuan Ni, Jian-Tao Sun, Jian Hu, Zheng (2011): Chen Cross Lingual Text Classification by Mining Multilingual Topics from Wikipedia. Proc. WSDM 2011.

[O'Connor 2012] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, Noah A. Smith (2010): From tweets to polls: Linking text sentiment to public opinion time series. Proc. International AAAI Conference on Weblogs and Social Media, 2010.

[Olsten 2010] Christopher Olston, Marc Najork (2010): Web Crawling. In: Foundations and Trends_ in Information Retrieval Vol. 4, No. 3 (2010) p. 175–246

[Och 2003] Franz Josef Och, Herman Ney (2003): A Systematic Comparison of Various Statistical Alignment Models. Journal Computational Linguistics, Vol. 29, 2003.

[Och 2004] Och, F. J., Ney, H. (2004): The alignment template approach to statistical machine translation. Computational Linguistics 30(4), 417-449, 2004.

[Osimo 2012] David Osimo, Francesco Mureddu, Riccardo Onori, Stefano Armenia (2012): International Research Roadmap on ICT Tools for Governance and Policy Modelling. Deliverable D2.2.1 of the Crossover project. URL: http://de.scribd.com/doc/116538462/D2-2-1-International-Research-Roadmap-on-ICT-Tools-for-Governance-and-Policy-Modelling-1

[Paass 2009] **Gerhard Paass, Frank Reichartz** (2009): Exploiting semantic constraints for estimating supersenses with CRFs. Proc. International Conference on Data Mining (SDM) 2009.

[Paass 2012] **Gerhard Paass** (2012): Document Classification, Information Retrieval, Text and Web Mining. In Alexander Mehler, Laurent Romary (eds.) Handbook of Technical Communication, pp. 141-189. de Gruyter, Berlin.

[Paass 2012a] **Gerhard Paass, André Bergholz, Anja Pilz** (2012): A knowledge-extraction approach to identify and present verbatim quotes in free text. Proc. I-KNOW 2012.

[Paolillo 2005] J. Paolillo, **D. Pimienta**, **D. Prado** (2005): Measuring Linguistic Diversity on the Internet, UNESCO, 12/2005.

[Paul 2011] Michael J. Paul, Mark Dredze (2011): You Are What You Tweet: Analyzing Twitter for Public Health. 5th Interational Conference on Weblogs and Social Media – 2011.

[Pilz 2011] **Anja Pilz, Gerhard Paass** (2011): From Names to Entities using Thematic Context Distance. Proceedings of 20th ACM Conference on Information and Knowledge Management (CIKM 2011).

[Pilz 2012] **Anja Pilz, Gerhard Paass** (2012): Collective Search for Concept Disambiguation. Int. Conf. on Computational Linguistics, Coling 2012.

[Pimienta 2009] **D. Pimienta, D. Prado**, **Á. Blanco** (2009): Twelve years of measuring linguistic diversity in the Internet: balance and perspectives, UNESCO, 2009

[Reichartz 2009] **Frank Reichartz, Hannes Korte, Gerhard Paass** (2009): Composite kernels for relation extraction. Proc. Association for Computational Linguistics ACL 2009.

[Reichartz 2010] **Frank Reichartz, Hannes Korte, Gerhard Paass** (2010): Semantic relation extraction with kernels over typed dependency trees. Proc. KDD 2010.

[Rubin 2012] Timothy Rubin, America Chambers, Padhraic Smyth, Mark Steyvers (2011): Statistical Topic Models for Multi-Label Document Classification. Machine Learning, Volume 88, pp 157-208.

[Singer 2012] E. Singer, P.A. Torres-Carrasquillo et al (2012): The MITLL NIST LRE 2011 Language Recognition System Odyssey 2012, Singapore, June 2012.

[Socher 2011] Richard Socher, Jeffrey Pennington, Eric H. Huang, Andrew Y. Ng, Christopher D. Manning (2011): Semi-Supervised Recursive Autoencoders for Predicting Sentiment Distributions. Proc. EMNLP '11.

[Sorg 2012] P. Sorg, P. Cimiano (2012): Exploiting Wikipedia for cross-lingual and multilingual information retrieval. Data & Knowledge Engineering, Volume 74, April 2012, pp. 26-45.

[Steeneken 1995] Herman JM Steeneken, David A. Van Leeuwen (1995): Multi-Lingual Assessment of Speaker. Independent Large Vocabulary Speech Recognition Systems: the SQALE Project. Eurospeech'95, Madrid.

[Suzuki 2002] Izumi Suzuki, Yoshiki Mikami, Ario Ohsato, Yoshihide Chubachi (2002): A Language and Character Set Determination Method Based on N-gram Statistics. in ACM Transactions on Asian Language Information Processing, Vol.1, No.3

[Sweeney 2002] L. Sweeney (2002): k-anonymity: A model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 557-570.

[Van Alstyne 2003] Van Alstyne, M., J. Zhang (2003) Emailnet: A system for automatically mining social networks from organizational email communication. Ann Arbor: Carnegie Mellon.

[Wahabzada 2010] M. Wahabzada, **Z. Xu, K. Kersting** (2010): Topic Models Conditioned on Relations. Proc. ECML PKDD.

[Wahabzada 2011] **M. Wahabzada, K. Kersting** (2011): Larger Residuals, Less Work: Active Document Scheduling for Latent Dirichlet Allocation. Proc. ECML PKDD 2011.

[Wallraf 2000] Wallraf, B. (2000): What global language". The Atlantic Monthly 286: 5 (2.000): pp. 52-66.

[Weber 1997] Weber, G. (1997): Top Languages". Language Today. December (1997): 12-18

[Wu 2011] Liang Wu, Yuanchun Zhou, Fei Tan, Fenglei Yang, Jianhui Li (2011): Generating Syntactic Tree Templates for Feature-Based Opinion Mining. Proc. ADMA 2011.

[Yamada 2001] Yamada, K. Knight, K. (2001): A syntax-based statistical translation model. In Proceedings of 39th Annual Meeting of the Association for Computational Linguistics (ACL 2001), 523-530, Toulouse, France, 2001.

[Zhang 2008] Zhang, M., Jiang, H., Aw, A., Li, H., Tan, C., Li, S. (2008): A Tree Sequence Alignment-based Tree-to-Tree Translation Model. In Proceedings of ACL-08: HLT, 559-567. Columbus, Ohio, USA, 2008.

[Zissman 1996] Marc A. Zissman (1996): Comparison of Four Approaches to Automatic Language Identification of Telephone Speech, IEEE Trans. Speech and Audio Processing, 4(1):31-44, 1996.

# B 5.2 Letters of Intent

## B 5.2.1 Observatoire Européen du Plurilinguisme


Paris, le 2 janvier 2013

Daniel Prado
Secrétaire général
de l'association MAAYA

We have been informed by MAAYA, one of the **SEMACORE** consortium partner, on their proposal for a call on the Program Framework 7 of European Commission and more precisely we received information on the task named **Linguistic Diversity Uses Cases**.

The **European Observatory for Plurilingualism** collects and analyses information from a network of partners. It publishes or refers to articles and existing studies. It produces its own original articles and can conduct studies. It draws up strategies that it proposes to those involved in the field of languages and plurilingualism.

The European observatory for plurilingualism is made up of a board of governors and a scientific council, and expresses its strong interest in the use of the data and expertise produced by the activity which purpose *is the weighting of languages in the main non national digital libraries*.

Additionally, we are eager to participate in any possibility of future exploitation in order to maintain some type of monitoring.

The integration of language measurement with the determination of contents in web documents is at the forefront of today's language diversity research and important for economic and cultural applications. The proposal submitted will advance the state of the art in this field.

Observatoire Européen
du Plurilinguisme
O. E. P.
4, rue Léon Séché
75015 PARIS - FRANCE

Christian TREMBLAY
Président de l'OEP

O E P
4 rue Léon Séché F-75015 Paris, France - http://www.observatoireplurilinguisme.eu/
Courriel : observatoire.plurilinguisme@wanadoo.fr - portable : 33 (0)6 10 38 68 90
SIRET : 502 404 866 00019

### B 5.2.2 Société Européenne de l'Internet

Société européenne de l'Internet
15, rue de l'Ancienne Comédie
75006 PARIS
http://www.ies-France.eu
+33 (0)6 63 24 39 14

Paris, le 6 janvier 2013

L'association Maaya nous a informé de sa participation au consortium européen SEMACORE qui va répondre à un appel à projets dans le cadre du Programme de Travail FP7 de la Commission européenne. Maaya nous a entretenus en particulier de la partie du projet intitulée **Linguistic Diversity Uses Cases**.

La Société européenne de l'Internet (SEI), association française sans but lucratif, est un groupe de réflexion qui réunit des experts de spécialités variées dans le domaine de l'Internet. Son objet principal est de contribuer au renforcement des capacités critiques de ses membres sur les questions concernant Internet et les réseaux de communication. Le groupe est également attentif à l'élaboration de cyberstrategies française et européenne, et en particulier à tout ce qui peut améliorer l'appropriation par le public des choix techniques et de leurs enjeux sociétaux.
La diversité linguistique et la mise à disposition, en ligne, de fonds culturels dans les différentes langues de l'Europe nous paraît un axe stratégique majeur. Disposer d'outils de compréhension du monde dans sa langue maternelle nous paraît essentiel à l'exercice de la citoyenneté.
La diversité linguistique est en outre un trait majeur de l'identité européenne. Elle constitue aussi un obstacle apparent susceptible d'être transformé en richesses à condition d'en explorer l'exploitation. Le projet Semacore nous paraît aller dans ce sens. L'appréciation de l'efficacité de politiques linguistiques doit en effet reposer sur une métrologie rigoureuse et continue.

En tant que présidente de la SEI, je suis particulièrement sensible à cet aspect du fait de ma formation initiale en lettres classiques et linguistique. J'agis par ailleurs comme analyste de l'évolution d'Internet sur le blog invité du journal *Le Monde* consacré aux réseaux : reseaux.blog.lemonde.fr.

La SEI est très intéressée à suivre les travaux de SEMACORE, à en examiner les résultats et à envisager les modalités de leur exploitation dans ses activités.

Avec nos meilleurs vœux pour 2013, veuillez recevoir l'expression de nos meilleures salutations.

Dominique Lacroix
présidente

### B 5.2.3 EuroLinc


Paris, January 5th, 2013

Dear Sir or Madam,

We have been informed by MAAYA, one of the **SEMACORE** consortium partner, on their proposal for a call on the Program Framework 7 of European Commission and more precisely we received information on the task named **Linguistic Diversity Uses Cases**.

**EUROLINC** is a non-profit dedicated on multilingualism in internet and natural language used by people in Europe, Africa, Asia, and expresses its strong interest in the activity which purpose is to characterize national web spaces and document guidelines to apply the method in other countries.

Furthermore, our institution is motivated to contribute in the exploitation of those methods in form of their application in other European countries

Expression of support to the project outcomes as for example: the measure of language diversity and the characterization of national Webs with innovative perspective will represent a sensible advance of the state of art in that field and open promising perspective of the application of linguistic diversity researches into the cultural and economic impact of the Internet.

Best Regards,

Chantal Lebrument,
Chairwoman

### B 5.2.4 African Network for Localisation

*African Network for Localisation*
192 Alemda Street, Queenswood, 0185, South Africa

To whom it may concern,

**Support for MAAYA in the SEMACORE application**

MAAYA, one of the **SEMACORE** consortium partner, has brought to our attention their proposal for Program Framework 7 of European Commission, more specifically the task **Linguistic Diversity Uses Cases**.

**ANLoc** is a network of African localisers, or software and technology translators. Our organisation focuses on advancing African languages in the digital age. We have run various projects across Africa focusing on removing barriers to African language software and digital content.

As director of Translate.org.za, a member organisation of ANLoc, I have personally helped to localise software into the 11 official languages of the country and grappled with the problem of stimulating content production and measuring the content production in the digital space.

We thus have a strong interest in the use of the data and expertise produced by the segment of **Linguistic Diversity Uses Cases** whose purpose *is to characterize national web spaces and document guidelines to apply the method in other countries*. For the simple reason that we would want to and need to implement the same approaches in our own countries.

**ANLoc** is ready to engage in future exploitation of such type of automatic observation and to integrate the results to help advance and focus our primary activities. With so many African languages we look forward to being able to firstly use such data to allow us to determine where to focus our next effort for maximum impact. But once we have completed our work we'd like to observe where our interventions lead to positive changes in the user community. We're excited that these types of questions will be more easily answered with the work that the **Linguistic Diversity Use** Cases produces.

Regards

Richard Dwayne Bailey
Research Director
African Network for Localisation
dwayne@translate.org.za
+27 12 460 1095

### B 5.2.5 Instituto Nacional de lenguas indigenas

SEP
SECRETARÍA DE EDUCACIÓN PÚBLICA

INALI
INSTITUTO NACIONAL DE LENGUAS INDÍGENAS

**DIRECCIÓN GENERAL**
**INALI.A.A.5.12/0001/2013**
**México, D.F., a 07 de enero 2013**

Daniel Prado
Executif Secretary
World Network for Linguistic Diversity

We have been informed that MAAYA, one of the SEMACORE consortium partner, on their proposal for a call on the Program Framework 7 of the European Commission, and more precisely we received information on the task names *Linguistic Diversity Uses Cases*.

Given that the National Institute of Indigenous Languages is the Mexican government's office that works for the preservation, strengthening and development of the linguistic diversity in Mexico, it is of their concern to express its strong interest in the use of data and expertise produced by the project mentioned above which purpose is to characterize national web spaces and document guidelines to apply the method in other countries.

INALI is ready to engage in future exploitation of such type of automatic observation and to integrate the results as an important tool on our daily activities.

Finally, the possibility to automatically get observation data about online linguistic diversity will help this Institute to improve its attention towards the indigenous peoples, as well as to develop new public policies within a multicultural framework.

Sincerely yours,

Javier López Sánchez
General Director

### B 5.2.6 Organisation Internationale de la Francophonie

ORGANISATION INTERNATIONALE DE la francophonie

***Direction de la langue française et de la diversité culturelle et linguistique***
*et*
***Direction de la Francophonie numérique***

N/Réf. : DLC/OBS/ITF/AW/gd/20121221-008

Dossier suivi par Alexandre Wolff
Téléphone : 01 44 37 33 85
Courriel : alexandre.wolf@francophonie.org

Paris, le 21 décembre 2012

**Objet :** Lettre d'intention en faveur du projet SEMACORE

Par la présente, l'OIF, représentée par son directeur de la Francophonie numérique, Monsieur Pierre Ouédraogo, et sa directrice a.i. de la langue française et de la diversité culturelle et linguistique, Madame Imma Tor Faus, confirme son intérêt et engagement pour le projet SEMACORE dans les termes ci-après.

L'Organisation internationale de la Francophonie est attentive à la place de la langue française et des nombreuses langues partenaires de ses 77 États membres ou observateurs dans le monde en général et dans le cyberespace en particulier. Elle dispose d'un Observatoire de la langue française qui publie régulièrement des données sur la présence de cette langue dans le monde. C'est dans ce contexte que l'OIF s'est intéressée à l'initiative de l'organisation MAAYA pour créer un consortium de recherches avancées susceptibles de dépasser les limites actuelles pour la production d'indicateurs des langues dans l'Internet, pour la caractérisation de celui-ci, l'analyse des contenus par langue et par pays, afin d'aider à la compréhension des évolutions de la structure de la Toile dont les contenus sont en évolution permanente.

À travers sa Direction de la Francophonie numérique et son Observatoire de la langue française, l'OIF participe financièrement aux pré-études conduites par MAAYA qui ont abouti à la définition du projet SEMACORE. L'OIF souhaite continuer à apporter sa contribution et participer au suivi de cette opération, notamment en proposant l'utilisation des résultats attendus de SEMACORE dans le but de mieux appréhender la place et l'utilisation de toutes les langues de la planète dans le monde numérique. Ce n'est qu'à partir de bons indicateurs, manquants à l'heure actuelle, que des politiques d'aménagement linguistique et d'alphabétisation numérique pourront se développer en vue de permettre à chaque citoyen de s'épanouir dans sa propre langue.

Conscients de l'importance des utilisations qui pourront être faites des produits issus des recherches liées au projet SEMACORE, nous acceptons de participer au Conseil de Surveillance que prévoit le projet sur les aspects sociétaux, avec des fonctions non rémunérées de conseil et d'assistance et dans un rôle d'évaluation de certains produits.

/ ...

Nous acceptons que cette lettre d'intention soit jointe à la proposition du projet SEMACORE et continuerons d'œuvrer pour le succès de cette démarche qui nous paraît déterminante pour le futur des indicateurs de la société de l'information et pour une meilleure caractérisation de la Toile.

**Pierre OUEDRAOGO**
Directeur de la Francophonie numérique

**Imma TOR FAUS**
Directrice a.i. de la langue française
et de la diversité culturelle et linguistique

### B 5.2.7 Associação Internacional Biblioteca Digital Lusófona (AIBDL)

Lisboa, 2013-01-13

**Associação Interncional Biblioteca Digital Lusófona (AIBDL)**

Praceta Estado da Baía, nº 12, R/C C, 2735-670,

S. Marcos, Cacém

Portugal

We have been informed by MAAYA, one of the SEMACORE consortium partner, on their proposal for a call on the Program Framework 7 of European Commission and more precisely we received information on the task named Linguistic Diversity Uses Cases.

The Associação Interncional Biblioteca Digital Lusófona aims to develop and establish methodologies, procedures for processing the digital cultural heritage of the Lusophone countries, as well as the defense and promotion of the Portuguese language in cyberspace, contributing to the knowledge, promotion and dissemination of the culture in the world.

From my experience studying and teaching Romanic Languages and Culture and Information Sciences (Ph.D) in different universities (Mozambique, Portugal and Spain), and as president of the association, and, although the Portuguese language is not one of the languages of the research, its strong interest in the use of the data and expertise produced by the activity which purpose is the weighting of languages in the main digital libraries.

Additionally, we are eager to participate in any possibility of future exploitation in order to maintain some type of monitoring.

The integration of language measurement with the determination of contents in web documents is at the forefront of today's language diversity research and important for economic and cultural applications. The proposal submitted will advance the state of the art in this field.

My respects

Presidente of the **Associação Internacional Biblioteca Digital Lusófona (AIBDL)**

Fernanda Maria Melo Alves

fmeloa2@hotmail.com