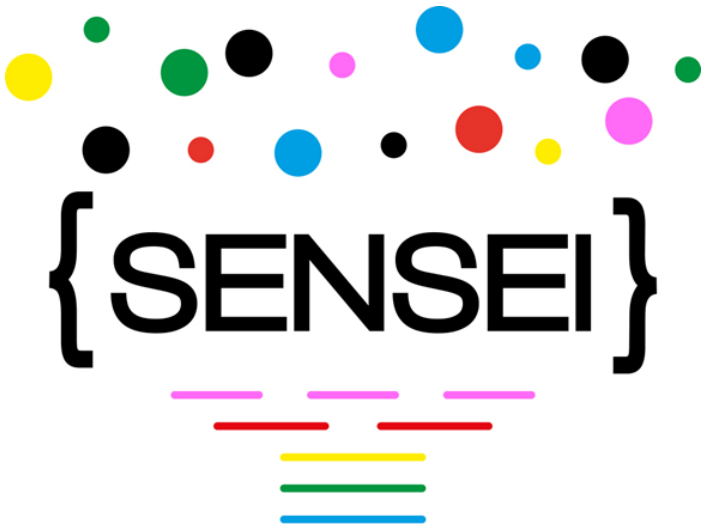

D8.4 – Second Ethical Issues Report

Document Number	D8.4
Document Title	Second Ethical Issues Report
Version	5.3
Status	Final
Work Package	WP8
Deliverable Type	Report
Contractual Date of Delivery	31.10.2015
Actual Date of Delivery	30.10.2015
Responsible Unit	UNITN
Keyword List	Ethical issues, data collection, user requirement processes
Dissemination level	PU

Editor

Giuseppe Riccardi (University of Trento, UNITN)

Contributors

Elisa Chiarani (University of Trento, UNITN)
Fabio Celli (University of Trento, UNITN)
Morena Danieli (University of Trento, UNITN)
Benoit Favre (University of Marseille, AMU)
Monica Lestari Paramita (University of Sheffield, USFD)
Vincenzo Giliberti (Teleperformance Italy, TP)
Letizia Molinari ((Teleperformance Italy, TP)
Hugo Zaragoza (Websays SL, Websays)
Marc Poch (Websays SL, Websays)

SENSEI Coordinator

Prof. Giuseppe Riccardi
Department of Information Engineering and Computer Science
University of Trento, Italy
giuseppe.riccardi@unitn.it

Document change record

Version	Date	Status	Author (Unit)	Description
0.1	20/07/2015	Draft	Elisa Chiarani, G. Riccardi (UNITN)	Table of Content and outline (who does what)
0.2	20/08/2015	Draft	Marc Poch (Websays)	Websays contribution
0.2	20/08/2015	Draft	Fabio Celli (UNITN)	UNITN contribution
0.3	24/08/2015	Draft	Monica Lestari Paramita (USFD)	USFD input
0.4	26/08/2015	Draft	Vincenzo Giliberti, Letizia Molinari (TP)	TP contributions
0.5	27/08/2015	Draft	Benoit Favre (AMU)	Add text about French data
1.0	27/08/2015	Draft	Elisa Chiarani (UNITN)	Merged initial draft
2.0	01/10/2015	Draft	Marc Poch (Websays), Hugo Zaragoza (Websays), Elisa Chiarani (UNITN)	Consolidated version
3.0	04/10/2015	Draft	Giuseppe Riccardi (UNITN)	Complete version
4.0	12/10/2015	Draft	Monica Lestari Paramita (USFD)	Review of v3.0
4.1	16/10/2015	Draft	Elisa Chiarani (UNITN)	Updated version after review
4.2	16/10/2015	Draft	Marc Poch (Websays)	Addition of agreement with Guardian and some comments fixed
5.0	19/10/2015	Draft	Elisa Chiarani (UNITN)	Updated version ready for final review
5.1	20/10/2015	Draft	Giuseppe Riccardi (UNITN)	Updated version ready for final review
5.2	22/10/2015	Final	Elisa Chiarani (UNITN)	Final version ready for submission
5.3	01/12/2015	Final	H. Zaragoza, M. Poch (Websays), G. Riccardi	Revised final version as requested at the Review

			(UNITN)	

Executive summary

In this document we report on the latest ethical issues mostly addressing the social media data collection for the technology development and annotations of speech summaries. In particular we have completed the process of categorizing the data sets in terms of their usage distribution requirements and rights.

Table of Content

1. Introduction

This document addresses the last issue in the pipeline of the data acquisition process. In particular, we have defined the guidelines we are going to use for the distribution of social media data. These guidelines allow us to pursue the future work in SENSEI with zero-risk strategy in terms of compliance with copyright laws without affecting the progress of SENSEI's technology development work. In fact, the main goal of SENSEI is to do technology development and evaluation and not data resource distributions. In the second part of this report we describe the last set of issues we addressed when collecting annotations of human summaries for both the speech and social media use cases.

1.1. Follow up - to P1 Reviewer Recommendations

Recommendation n. 7: *Ethical issues with user data from social media and newspaper should be addressed*.

All SENSEI data sources have been assigned a Data Usage Level (DUL) code based on their category of terms and conditions. Letters asking for permission have been sent to data resource owners for which a written approval was required to redistribute their data. For those who responded and gave the SENSEI consortium a written approval, their data will be used accordingly. On the other hand, those sources for which there has been no response or a negative one those sources will be reassigned the lowest DUL, i.e. DUL3: "URLs Only" (only URLs will be distributed).

2. Social Media and IPR in Data Sources

In this section we are going to present the different Intellectual Property Rights (IPR) scenarios found in the SENSEI data sources and how the SENSEI consortium is addressing those scenarios to work with such data.

Other EU projects have faced similar situations and needs regarding IPR.

The approach followed by PANACEA project, a relevant FP7-funded project which aimed at building a factory of language resources, is presented in Section 2.1 and has been taken into account when making this survey.

Terms and conditions of microblogging sites and online newspapers have then been surveyed in order to obtain a global vision of the necessary IPR to use, modify and distribute that data.

To simplify and facilitate the analysis, IPR scenarios have been simplified to four different levels of usage: “Full Data”, “Data Under Agreement (for content and UGC), “Data Under Agreement (for content only)” and “URLs Only” as shown in Table 1 in Section 2.2.

In Section 0, we reviewed terms and conditions of all SENSEI data sources and sought approvals to use and redistribute their data where appropriate. These data usage levels (both with and without approval) are summarised in Table 2 in Section 2.4.

This document also includes the final table (Table 3 in Section 2.5) listing all data sources and their assigned DULs which took into account the responses to the letters that have been sent.

2.1. PANACEA project’s approaches

In this section we report the approach of another European project, PANACEA, which has attempted to address these issues in the past.

PANACEA[1] project goal was the creation of language resources. The main approach for gathering data for those language resources was crawling the web.

PANACEA’s deliverable D2.4: "Licensing Policy and Exploitation Plan"[2] explains how these issues were faced in the project. Section 3.3 talks about language resources.

Extracted from that deliverable:

> …
>
> *The paradigmatic case would be web crawling for deriving any type of Language Resource. This is an important facet of the PANACEA Platform exploitation and we would like to elaborate more on the various issues involved. Data crawling may be defined as the act of collecting different forms of information from the public Internet in an automatic fashion, which is then stored and processed in different ways. Crawling could be the initial step for different PANACEA web service and workflow operations. Text would be automatically compiled from different, automatically selected web sites. The first implication is that some of the crawled data may be protected e.g. for*

[1] EC FP7 project PANACEA (http://www.panacea-lr.eu), Grant Agreement 248064

[2] http://www.panacea-lr.eu/system/deliverables/PANACEA_D2.4.pdf

copyright considerations. However, not all texts in the web are protected. For instance, works made by the public sector agencies remain outside the protection of copyright law for most European legislation jurisdictions. The second implication has to do with what are the acts that are going to be performed upon the data, and then to assess two factors: (a) the degree to which results of such acts fall within the acts restricted by laws and (b) the extent to which results of such acts are visible enough to expose a PANACEA user to the risk of legal action. PANACEA has got expert advice about these two aspects. The main rationales behind this and the conclusions are important to our exploitation plan and hence are fully appended to this report. A brief summary of the recommendations is given herein:

(a) In the case of web crawling for LT purposes, the acts indeed include copying and processing of the texts and also the creation of derivative works with the aim of being used and/or communicated to third parties. These acts require separate permission of the right holders, unless this is clearly indicated in a legal statement or the site indicates being offered under some permissive licenses such as Creative Commons.

(b) Though the act of web crawling is part of the daily operation of a web site and could be covered by an implicit license, the owner of a web site has technological means to prevent the site to be indexed or copied. Thus, if they do not exist, it might imply that the web site owner wishes it to be copied. To rely on the implied license might be considered a rather prudent attitude even though the goal of crawling for LT purposes may not have been the intention of the web site owner. However, legal advice suggests to reduce the risks crawling for LT, by doing the following:

(i) only crawl sites where bots are allowed

(ii) include a public notice stating that their works only derive from web sites that do not prohibit crawling;

(iii) include a notice and a takedown procedure indicating under which circumstances the material will be taken down, what the decision making procedure is and an e-mail address where relevant complaints could be addressed.

…

Summary: Unless indicated in Creative Commons or similar licenses, pemission needs to be asked from the owner of each crawled webpage. There was no attempt to contact authors of UGC, instead the project relied on the owner's right to grant use of the data.

2.2. Data Use Levels (DULs)

Four data use levels have been created to identify the varying degrees of data sharing – both for the content and the user-generated contents (UGC) – allowed by each data source.

Table 1. Data Use Levels (DULs)

Data Use Levels (DULs)	Description
DUL 0: Full Data	We can use, edit and share the **Content** and the **UGC** ***without written approval***

DUL 1: Data Under Agreement	We can use, edit and share the **Content** and the **UGC *with written approval***
DUL 2: Data Under Agreement	We can use, edit and share the **Content *with written approval***
DUL 3: URLs Only	We can share the URLs, but not the content and the UGC.

Distributing URLs instead of Collections

It is often the case that content is accesible publicly (A and B can download the content) but protected from distribution (A can not send the content to B). This is the case for example for most WWW content, including web pages, news pages and Social Media content. Since all content in the web is accessed through URLs, which by definition is public and can be shared, any web collection can be represented by a list of URLs, which can be distributed without risk. From this list anyone interested can download the content and reconstruct the collection (A creates a collection with three pages {P1, P2 and P3} with URLS {U1, U2 and U3}. Instead of sending the collection to B, she send the list of urls {U1,U2,U3} to B and B downloads the content, effectively reconstructing the same collection. This technique has been extensively used for research, for exaple in international competitions such as TREC (US) and SemEval (EU).

2.3. Data sources

We listed all data sources which we explored in this project and reported their relevant contents in their terms and conditions, followed by the assigned DULs for their data.

2.3.1. Microblogging sites

2.3.1.1. Twitter

Twitter Basic Terms:

> *You are responsible for your use of the Services, for any Content you post to the Services, and for any consequences thereof. Most Content you submit, post, or display through the Twitter Services is public by default and will be able to be viewed by other users and through third party services and websites. You should only provide Content that you are comfortable sharing with others under these Terms.*
>
> ***UGC**: You retain your rights to any Content you submit, post or display on or through the Services. By submitting, posting or displaying Content on or through the Services, you grant us a worldwide, non-exclusive, royalty-free license (with the right to sublicense) to use, copy, reproduce, process, adapt, modify, publish, transmit, display and distribute such Content in any and all media or distribution methods (now known or later developed).*
>
> ***Tip**: This license is you authorizing us to make your Tweets on the Twitter Services available to the rest of the world and to let others do the same.*
>
> *Except as permitted through the Twitter Services, these Terms, or the terms provided on dev.twitter.com, you have to use the Twitter API if you want to reproduce, modify, create derivative works, distribute, sell, transfer, publicly display, publicly perform, transmit, or otherwise use the Twitter Services or Content on the Twitter Services.*

***Tip**: We encourage and permit broad re-use of Content on the Twitter Services. The Twitter API exists to enable this.*

Summary: Twitter encourages a broad re-use of their data but always making use of the Twiter API. Although in theory a tweet could be copyrighted, 99% of them could not due to their short size [1] therefore we should be able to download, edit, distribute, etc. However there is always a small risk of infringing IPR. For this reason, for SENSEI project we will treat Twitter content as DUL 3.

Rights: **DUL3: URLs only**: we can share the URLs but not the content.

2.3.1.2. Facebook

Facebook Basic Terms:

Statement of Rights and Responsibilities

This Statement of Rights and Responsibilities ("Statement," "Terms," or "SRR") derives from the Facebook Principles, and is our terms of service that governs our relationship with users and others who interact with Facebook, as well as Facebook brands, products and services, which we call the "Facebook Services" or "Services". By using or accessing the Facebook Services, you agree to this Statement, as updated from time to time in accordance with Section 13 below. Additionally, you will find resources at the end of this document that help you understand how Facebook works.

Because Facebook provides a wide range of Services, we may ask you to review and accept supplemental terms that apply to your interaction with a specific app, product, or service. To the extent those supplemental terms conflict with this SRR, the supplemental terms associated with the app, product, or service govern with respect to your use of such app, product or service to the extent of the conflict.

Privacy

Your privacy is very important to us. We designed our Data Policy to make important disclosures about how you can use Facebook to share with others and how we collect and can use your content and information. We encourage you to read the Data Policy, and to use it to help you make informed decisions.

Sharing Your Content and Information

You own all of the content and information you post on Facebook, and you can control how it is shared through your privacy and application settings. In addition:

For content that is covered by intellectual property rights, like photos and videos (IP content), you specifically give us the following permission, subject to your privacy and application settings: you grant us a non-exclusive, transferable, sub-licensable, royalty-free, worldwide license to use any IP content that you post on or in connection with Facebook (IP License). This IP License ends when you delete your IP content or your account unless your content has been shared with others, and they have not deleted it.

When you delete IP content, it is deleted in a manner similar to emptying the recycle bin on a computer. However, you understand that removed content may persist in backup copies for a reasonable period of time (but will not be available to others).

When you use an application, the application may ask for your permission to access your content and information as well as content and information that others have shared with you. We require applications to respect your privacy, and your agreement with that application will control how the application can use, store, and transfer that content and information. (To learn more about Platform, including how you can control what information other people may share with applications, read our Data Policy and Platform Page.)

When you publish content or information using the Public setting, it means that you are allowing everyone, including people off of Facebook, to access and use that information, and to associate it with you (i.e., your name and profile picture).

We always appreciate your feedback or other suggestions about Facebook, but you understand that we may use your feedback or suggestions without any obligation to compensate you for them (just as you have no obligation to offer them).

Summary: User generated content (including content covered by IP rights) remains property of the user. A licence to share and distribute is grant to Facebook by the User when shared, used, etc. in Facebook. The content can be accessed by anyone from the Facebook page If it is shared with the Public setting. For this reasons we we consider Facebook content as DUL3.

Rights: DUL3: urls only: we can share the urls but not the content.

2.3.1.3. Microblogging and other Social Media sites summary

After surveying a few other microblogging sites (Twitter and Facebook presented in this section) and analyzing their Terms and Conditions we can conclude that most of them present content which could be shared and distributed without a big risk of being sued because it has been made public by users. However, most microblogging sites claim that the user remains the IP owner of the shared content. If the user is the IP rights owner then it is almost impossible to ask all of them for permission instead of only getting permission from the sites (as usually happens with newspapers). Therefore, as a safe position, SENSEI will consider all microblogging sites as DUL3, and will only share or distribute links to the content.

2.3.2. Online newspaper

In this section, we describe the terms of services of different newspapers (as of January 2015) and discuss their implications.

2.3.2.1. The Guardian

URL: http://www.theguardian.com/

Terms of Service: http://www.theguardian.com/help/terms-of-service

> ***3. Use of material appearing on the Guardian Site***
>
> *You may download and print extracts from the Guardian Content* ***for your own personal and non-commercial use only****, provided you maintain and abide by any author attribution, copyright or trademark notice or restriction in any material that you download or print. You may not use any Guardian Content for any other purpose without our* ***prior written approval****. Except as expressly authorised by the Guardian,* ***you are not allowed to create a database in electronic or paper form comprising all or part of the material appearing on the Guardian Site****.*
>
> ***6. User content***
>
> *You or the owner of the content still own the copyright in the content sent to us, but by submitting content to us, you are granting us an unconditional, irrevocable, non-exclusive, royalty-free, fully transferable, perpetual worldwide licence to use, publish and/or transmit, and to* ***authorise third-parties to use, publish and/or transmit your content*** *in any format and on any platform, either now known or hereinafter invented.*

Summary: With written approval we can use the paper content and the user-generated content.

Rights: SENSEI has an agreement with The Guardian: therefore **DUL1: Data under agreement**.

2.3.2.2. The Independent

URL: http://www.independent.co.uk/

Terms of Service: http://www.independent.co.uk/service/user-policies-759573.html

> ***Your promise to us***
>
> *The Website and the Content may only be used for your personal, non-commercial use.*
>
> *For this purpose alone you may retrieve and display the Content on a computer screen. You may also print out, but not photocopy, one copy of individual files on paper and store files in electronic form on disc, but not on any server or any other storage device connected to a network where the Content could be accessed by other users.*
>
> *Except as set out above, you agree not to download, copy, reproduce, modify, store, archive, show in public, redistribute or commercially exploit in any way any part of the Content* ***without the prior written permission*** *of Independent.co.uk.*
>
> ***User generated content***

In relation to UGC, when you submit any Contribution to the Website, whether that includes text, photographs, graphics, video or audio in any format, you agree by submitting your Contribution to grant the Company a perpetual, irrevocable, royalty-free, non-exclusive sub licensable right and licence to use, reproduce, modify, adapt, publish, translate, with respect to your Contribution worldwide and/or to incorporate your Contribution in other works in any media now known or later developed to the full term of any rights that may exist in your Contribution.

Summary: With written approval, we can use the paper content and the UGC.

Rights: Since we have no agreement, this source is assigned to **DUL3: URLs Only**.

2.3.2.3. The Standard

URL: http://www.standard.co.uk/

Terms of Service: http://www.standard.co.uk/service/terms-of-use-6902768.html

Your promise to us

The Website and the Content may only be used for your personal, non-commercial use.

For this purpose alone you may retrieve and display the Content on a computer screen. You may also print out, but not photocopy, one copy of individual files on paper and store files in electronic form on disc, but not on any server or any other storage device connected to a network where the Content could be accessed by other users.

Except as set out above, you agree not to download, copy, reproduce, modify, store, archive, show in public, redistribute or commercially exploit in any way any part of the Content without the prior written permission of Standard.co.uk

User generated content

In relation to UGC, when you submit any Contribution to the Website, whether that includes text, photographs, graphics, video or audio in any format, you agree by submitting your Contribution to grant the Company a perpetual, irrevocable, royalty-free, non-exclusive sub licensable right and licence to use, reproduce, modify, adapt, publish, translate, with respect to your Contribution worldwide and/or to incorporate your Contribution in other works in any media now known or later developed to the full term of any rights that may exist in your Contribution.

Summary: We can use the paper content and the UGC.

Rights: Since we have no agreement, this source has been assigned to **DUL3: URLs Only**.

2.3.2.4. El País

URL: http://elpais.com

Terms of Service: http://elpais.com/estaticos/aviso-legal/

Propiedad Intelectual e Industrial

...quedan expresamente prohibidas la reproducción, la distribución y la comunicación pública, incluida su modalidad de puesta a disposición, de la totalidad o parte de los contenidos de esta página web, con fines comerciales, en cualquier soporte y por cualquier medio técnico, sin la autorización de EDICIONES EL PAIS. El

USUARIO se compromete a respetar los derechos de Propiedad Intelectual e Industrial titularidad de EDICIONES EL PAIS. Podrá visualizar los elementos del Site e incluso imprimirlos, copiarlos y almacenarlos en el disco duro de su ordenador o en cualquier otro soporte físico siempre y cuando sea, única y exclusivamente, para su uso personal y privado.

Summary: With written approval we can use the paper content. There was no discussion regarding the UGC property rights.

Rights: Two emails were sent asking for a written approval but no answer was received so far. Since no response was received, this data source is assigned to **DUL3: URLs Only**.

2.3.2.5. El Mundo

URL: http://www.elmundo.es/

Terms of Service: http://www.elmundo.es/registro/avisolegal.html

Asimismo está prohibido modificar, copiar, reutilizar, explotar, reproducir, comunicar públicamente, hacer segundas o posteriores publicaciones, cargar archivos, enviar por correo, transmitir, usar, tratar o distribuir de cualquier forma la totalidad o parte de los contenidos incluidos en el Sitio Web si no se cuenta con la autorización expresa y por escrito de UNIDAD EDITORIAL o, en su caso, del titular o titulares de los derechos a que corresponda.

En caso de que esté interesado en una autorización o licencia para utilizar en cualquier forma los contenidos del Sitio Web diríjase a http://www.elmundo.es/elmundo/formularios/generico.html.

El Usuario cede a título gratuito a UNIDAD EDITORIAL, sin carácter de exclusiva, los derechos de propiedad industrial e intelectual y cualquier otro tipo de derecho sobre los Contenidos Generados, incluyendo, sin limitación, los derechos de reproducción, distribución, comunicación pública (incluida su modalidad de puesta a disposición) y transformación para su explotación por cualquier medio, soporte o formato y a través de cualquier sistema, procedimiento o modalidad de transmisión, comunicación o distribución, ya sea gratuito o de pago, utilizando los Contenidos Generales solos o conjuntamente con otro. La cesión a favor de UNIDAD EDITORIAL se realiza para todo el mundo, por todo el tiempo de duración de los derechos y comprende la facultad de cesión a terceros.

Summary: With written approval, we can use the paper content and the UGC.

Rights: An email asking for a written approval was sent but no answer has been received so far. Since no response was received, this data source is assigned to **DUL3: URLs Only**.

2.3.2.6. El Periódico

URL: http://www.elperiodico.com

Terms of Service: http://www.elperiodico.com/es/avisolegal.shtml

> *...quedando, por tanto, terminantemente prohibida la utilización de la totalidad o parte de los contenidos de este web y páginas webs pertenecientes al mismo con propósitos públicos o comerciales, su distribución, comunicación pública, incluida la modalidad de puesta a disposición, así como su modificación, alteración o descompilación a no ser que para ello se cuente con el consentimiento expreso y por escrito del titular del web.*
>
> *El usuario cede a título gratuito a GRUPO ZETA, sin carácter de exclusiva, los derechos de reproducción, distribución, transformación y comunicación pública, en todas las posibles modalidades, en relación a los contenidos enviados (fotografías, imágenes con o sin movimiento, textos, informaciones, bases de datos, grabaciones sonoras o cualesquiera otras obras o prestaciones). La citada cesión se realiza para todo el mundo, por la duración máxima actualmente prevista en la Ley de Propiedad Intelectual y con expresa facultad de cesión a terceros.*

Summary: With written approval we can use the paper content and the UGC.

Rights: An email asking for a written approval was sent but no answer was received so far. Therefore, this data source is assigned to **DUL3: URLs Only**.

2.3.2.7. La Provence

URL: http://www.laprovence.com

Terms of Service: http://www.laprovence.com/cgu

La Provence can modify and redistribute UGC:

> *...concède à la Société et au Groupe le droit d'utiliser, de reproduire, publier et diffuser les Contributions qu'il a décidé de stocker, transmettre ou mettre en ligne sur les Services Communautaires du Site, aux fins de fourniture des services d'information, et/ou à des fins promotionnelles, et ce, sur tout support électromagnétique et papier gratuit ou payant et par tout moyen de communication électronique et de presse, exploité par la Société ou par toute société du Groupe...*

Other users of the site can modify and redistribute UGC, under **applicable** law:

> *Compte tenu du caractère interactif du réseau internet, le Membre est averti que sa Contribution pourra être accessible, reproduite ou exploité par les Utilisateurs du Site... tout Utilisateur s'engage dans tous les cas à utiliser et exploiter la Contribution publiée par les Membres conformément à la réglementation applicable et notamment dans le respect des droits des tiers et du Membre auteur de la Contribution.*

Summary: The ToS describes that the company owns redistribution and modification rights on UGC as do users of the site, however, it does not clearly describe theuse by a third party. What is not clear is the passage about applicable law which layers should look at. It seems that under applicable law (French copyright law), we don't need an agreement if we reproduce UGC without distributing it. However, we do need agreement from the original author for redistribution, unless La Provence grants us this right.

Rights: Since we have no agreement, this source is assigned to **DUL3: URLs Only**.

2.3.2.8. Le Figaro

URL: http://www.lefigaro.fr

Terms of Service: http://mentions-legales.lefigaro.fr/page/cgu

Explicitly disallow reproduction of content (but exemptions can be negotiated, for example for RSS).

> *L'Utilisateur s'interdit de reproduire et/ou d'utiliser les marques et logos présents sur le Site, ainsi que de modifier, copier, traduire, reproduire, vendre, publier, exploiter et diffuser dans un format numérique ou autre, tout ou partie des informations, textes, photos, images, vidéos et données présents sur le Site, qui constituent des œuvres au sens des dispositions de l'article L112-1 du code de la propriété intellectuelle.*

UGC can be reproduced and formatted by Figaro under "indexing" rules.

> *L'Utilisateur, en déposant sa contribution, accepte sans contrepartie que celle-ci soit reproduite en tout ou partie, selon des règles d'indexation et de mise en forme qui sont du seul ressort du Figaro.. Toutefois, Le Figaro n'est pas responsable du contenu des contributions en général, en ce inclus les contributions de l'Utilisateur.*

Summary: Le Figaro content cannot be redistributed (might include UGC), however, there are no specificities on UGC use by third parties. French Law is applicable.

Rights: Since we have no agreement, this source has been assigned to **DUL3: URLs Only**.

2.3.2.9. Le Monde

URL: http://www.lemonde.fr/

Terms of Service: http://www.lemonde.fr/service/mentions_legales.html

No distribution of Le Monde content:

> *Ce droit est consenti dans le cadre d'un usage strictement personnel, privé et non collectif, toute mise en réseau, toute rediffusion ou commercialisation totale ou partielle de ce contenu, auprès des tiers, sous quelque forme que ce soit, étant strictement interdite.*

More information:

http://www.lemonde.fr/services-aux-internautes/article_interactif/2005/03/16/faq-droit-et-autorisation_626149_3388.html

Summary: Redistribution of Le Monde content is disallowed. There is no explicit mention of UGC, so it should be considered as Le Monde content. It is mentioned that Le Monde content can be cited, so referring to the data in research papers may be allowed. Note that Le Monde URLs are modified when an article is placed behind the paywall (after a period of time).

Rights: **DUL3: URLs Only**.

2.3.2.10. L'express

URL: http://www.lexpress.fr/

Terms of Service: http://www.lexpress.fr/outils/conditions-generales-utilisation.asp

L'Utilisateur accepte que ses contributions dans le cadre des Espaces de dialogue deviennent des informations publiques. L'Utilisateur accepte que ses contributions soient publiées, reproduites, modifiées, traduites, distribuées, présentées et/ou affichées, seules ou associées à d'autres travaux, sous toute forme, tout support ou toute technologie, actuellement connus ou inconnus. L'Utilisateur concède aux autres Utilisateurs, le droit d'accéder, afficher, enregistrer et reproduire les communications pour leur usage personnel, Groupe Express-Roularta étant dégagé de toute responsabilité à cet égard.

Summary: L'express allows full reuse, modification, redistribution of UGC to "other users" of the site.

Rights: A written approval was obtained on August 27th, 2015. **DUL1: Data Under Agreement**.

2.3.2.11. Les echos

URL: http://www.lesechos.fr/

Terms of Service: http://www.lesechos.fr/pratique/cgu.htm

utiliser et/ou télécharger les Informations sur son équipement que pour un usage exclusivement personnel, non marchand et limité dans le temps ; n'imprimer les Informations téléchargées sur support papier qu'à la condition que les copies ainsi constituées fassent l'objet d'un usage exclusivement personnel, ce qui exclut notamment toute reproduction à des fins professionnelles ou commerciales ou de diffusion en nombre, gratuite ou payante; ne pas conserver les Informations téléchargées plus de quarante-huit (48) heures et procéder à leur destruction passé ce délai ; à ne pas recopier tout ou partie du Site sur un autre site ou un réseau interne d'entreprise; ne pas extraire ou réutiliser, y compris à des fins privées, sans autorisation écrite et préalable des ECHOS, une partie substantielle ou non du contenu des bases de données et archives constituées par le Site; LES ECHOS est le propriétaire exclusif de tous les droits de propriété intellectuelle portant tant sur la structure que sur le contenu du Site.

http://www.lesechos.fr/pratique/cgu-espaces-de-dialogues.htm

Enfin, le contenu des Contributions, rendu anonyme, pourra être communiqué, à titre exceptionnel, à un organisme de recherche ou à un centre universitaire, à des fins de recherche.

Summary: No storage of downloaded data allowed, no extraction/reuse of the data even in a private setting (without written consent). However, there is a small paragraph of interest: anonymized UGC might be used for research, exceptionally. It is not clear if an agreement with Les Echos is necessary for such use.

Rights: Since we have no agreement this source is assigned to **DUL3: URLs Only**.

2.3.2.12. Libération

URL: http://www.liberation.fr/

Terms of Service: http://www.liberation.fr/licence/ http://www.liberation.fr/cguv/

Toute mise en réseau, toute rediffusion, sous quelque forme que ce soit, partielle ou totale, sont donc explicitement interdites sans l'accord exprès et préalable de la Sarl Libération.

Summary: No mention to UGC, not allowed to reproduce content of the website.

Rights: Since we have no agreement, this source is assigned to **DUL3: URLs Only**.

2.3.2.13. 20 Minutes

URL: http://www.20minutes.fr/

Terms of Service: http://www.20minutes.fr/cgu.php

> *Par ailleurs, en utilisant l'un ou l'autre des Services Interactifs objet des conditions générales d'utilisation, l'Inscrit est présumé irréfragablement avoir acceptéque tout Contenu dont il est à l'origine soit reproduit par 20 Minutes, tout autre Inscrit et les internautes accédant au site 20minutes.fr sans limitation de durée et aux seuls fins de création, de gestion et d'administration des services interactifs par 20 Minutes ou d'utilisation d'un ou plusieurs Services Interactifs ou de simples consultations du site 20minutes.fr.*
>
> *A ce titre, il reconnaît qu'il n'est cessionnaire d'aucun autre droit que celui d'utiliser les moyens notamment techniques mis à sa disposition par 20 Minutes pour utiliser un plusieurs des Services Interactifs. L'Inscrit est informé qu'il n'est pas autorisé à procéder au "reverse ingeneering" de tout ou partie des Services Interactifs et notamment du logiciel mis en oeuvre, ni à compiler ou désassembler ledit logiciel, ni à modifier ou compléter celui-ci ou tout ou partie des Services Interactifs proposés.*

Summary: 20 Minutes allows redistribution by of UGC by other registered users. Reverse-engineering of services is disallowed. No specific mention to the use of content by unregistered users, so French copyright law is applicable.

Rights: Since we have no agreement, this source is assigned to **DUL3: URLs Only**.

2.3.2.14. Metronews (FR)

URL: http://www.metronews.fr/

Terms of Service: http://www.metronews.fr/mentions-legales/cgu.xml

> *Vous cédez à METRO à titre gratuit, ainsi qu'à toutes sociétés du groupe auquel elle appartient, l'intégralité de vos droits d'auteur relatifs à vos contenus Contributifs, pour le monde entier et pour une durée de cinq ans à compter de la date de leur insertion sur le Site par METRO aux fins de les voir reproduire, représenter en public et diffuser sur le Site ainsi que sur tous autres supports et par tous autres moyens existants ou à venir, en totalité ou par extraits.*
>
> *Tous désassemblages, décompilations, décryptages, extractions, réutilisations, copies et plus généralement, tous actes de reproduction, représentation, diffusion et utilisation de l'un quelconque de ses éléments, en tout ou partie, sans l'autorisation de METRO, sont strictement interdits et feront l'objet de poursuites judiciaires.*

Summary: Metronews owns all UGC. Any redistribution of content is illegal.

Rights: Since we have no agreement this source is assigned to **DUL3: URLs Only**.

2.3.2.15. Corriere della Sera

URL: http://www.corriere.it/

Terms of Service: http://www.corriere.it/privacy.shtml

Comunichiamo a tutti gli interessati che a far data dal 1 gennaio 2012 il Titolare del Trattamento dei dati personali raccolti su questo sito ai sensi e per gli effetti del Codice della Privacy è la società RCS MediaGroup S.p.A. con sede in Milano via Rizzoli 8; l'esercizio dei diritti di cui all'art. 7 del D.Lgs 196/03 potrà effettuarsi attraverso specifica comunicazione a mezzo posta indirizzata alla medesima Società. Vi invitiamo a leggere il testo dell'informativa resa ai sensi dell'art. 13 del Codice della Privacy qui di seguito.

Conformemente all'impegno e alla cura che RCS MediaGroup S.p.A dedicano alla tutela dei dati personali, La informiamo sulle modalità, finalità e ambito di comunicazione e diffusione dei Suoi dati personali e sui Suoi diritti, in conformità all'art. 13 del D. Lgs. 196/2003.

Per offrirLe i servizi personalizzati previsti dai nostri siti internet, RCS MediaGroup S.p.A in qualità di Titolare del trattamento e deve trattare alcuni dati identificativi necessari per l' erogazione del Servizio.

Dati di navigazione

Le procedure software e il sistema informatico preposto al funzionamento dei siti web acquisiscono, nel corso del loro normale esercizio, alcuni dati la cui trasmissione è implicita nell'uso dei protocolli di comunicazione di Internet.

Queste informazioni non sono raccolte per essere associate a interessati identificati, ma che per loro stessa natura potrebbero, attraverso elaborazioni ed associazioni con dati detenuti dal Titolare o da terzi, permettere di identificare gli utenti.

In questa categoria di dati rientrano gli indirizzi IP o i nomi a dominio dei computer utilizzati dagli utenti che si connettono al sito, gli indirizzi in notazione URI (Uniform Resource Identifier) delle risorse richieste, l'orario della richiesta, il metodo utilizzato nel sottoporre la richiesta al server, la dimensione del file ottenuto in risposta, il codice numerico indicante lo stato della risposta data dal server (buon fine, errore, ecc.) ed altri parametri relativi al sistema operativo e all'ambiente informatico dell'utente.

Questi dati potranno essere utilizzati dal Titolare al solo fine di ricavare informazioni statistiche anonime sull'uso del sito al fine di individuare le pagine preferite dagli utenti in modo da fornire contenuti sempre più adeguati e per controllarne il corretto funzionamento. I dati potrebbero essere utilizzati per l'accertamento di responsabilità in caso di ipotetici reati informatici ai danni del sito.

Cookies

I cookies sono dei files che possono essere registrati sul disco rigido del suo computer. Questo permette una navigazione più agevole e una maggiore facilità d'uso del sito stesso.

I cookies possono essere usati per determinare se è già stata effettuata una connessione fra il suo computer e le nostre pagine. Viene identificato solo il cookie memorizzato sul suo computer.

Naturalmente è possibile visitare il sito anche senza i cookies. La maggior parte dei browser accetta cookies automaticamente. Si può evitare la registrazione automatica dei cookies selezionando l'opzione "non accettare i cookies" fra quelle proposte. Per avere ulteriori informazioni su come effettuare questa operazione si può fare riferimento alle istruzioni del browser. E' possibile cancellare in ogni momento eventuali cookies già presenti sul disco rigido. La scelta di non far accettare cookies dal browser può limitare le funzioni accessibili sul nostro sito.

Dati personali

I dati personali che Lei fornirà verranno registrati e conservati su supporti elettronici protetti e trattati con adeguate misure di sicurezza anche associandoli ed integrandoli con altri DataBase.

I dati e i cookies da Lei ricevuti saranno trattati da RCS MediaGroup S.p.A esclusivamente con modalità e procedure necessarie per fornirLe i servizi da Lei richiesti.

I dati non saranno diffusi ma potranno essere comunicati, ove necessario per l'erogazione del servizio, alle Società del Gruppo RCS Mediagroup, oltre che a società che svolgono per nostro conto compiti di natura tecnica od organizzativa strumentali alla fornitura dei servizi richiesti.

Solo con il Suo espresso consenso i dati potranno essere utilizzati per effettuare analisi statistiche, indagini di mercato e invio di informazioni commerciali sui prodotti e sulle iniziative promozionali di RCS MediaGroup S.p.A e/o di società terze.

Inoltre, sempre con il Suo consenso esplicito, tali dati potranno essere forniti ad altre Aziende operanti nei settori editoriale, finanziario, assicurativo automobilistico, largo consumo, organizzazioni umanitarie e benefiche le quali potranno contattarLa come Titolari di autonome iniziative - l'elenco aggiornato è a Sua disposizione e può essere richiesto al Responsabile del trattamento all'indirizzo sottoriportato - per analisi statistiche, indagini di mercato e invio di informazioni commerciali sui prodotti e iniziative promozionali.

Successivamente alla registrazione necessaria per il Servizio richiesto, ove Lei intenda richiedere/usufruire di altri Servizi erogati dalla stessa società o da società diverse come sopra indicate (RCS MediaGroup S.p.A o società da queste controllate o collegate) Lei potrà utilizzare le credenziali (nome utente/ mail/ password) già utilizzate per la prima registrazione.

Ove necessario Le potranno essere richiesti dati aggiuntivi, necessari per l'erogazione degli ulteriori servizi richiesti.

In ogni momento Lei potrà rileggere l'informativa ed eventualmente modificare i consensi precedentemente forniti, verificare e/o modificare lo stato dei servizi attivi ed eventualmente richiedere servizi aggiuntivi.

Il conferimento dei dati è facoltativo, salvo per quelli indicati come obbligatori per poterle permettere di accedere ai servizi offerti. Lei ha diritto di conoscere, in ogni momento, quali sono i Suoi dati e come essi sono utilizzati. Ha anche il diritto di farli aggiornare, integrare, rettificare o cancellare, chiederne il blocco ed opporsi al loro trattamento. Ricordiamo che questi diritti sono previsti dal Art.7 del D. Lgs 196/2003.

L' elenco aggiornato dei Responsabili del Trattamento dati di cui alla presente informativa è consultabile presso la Sede legale di RCS MediaGroup S.p.A in via A.Rizzoli, 8 - 20132 Milano; l'esercizio dei diritti di cui all'art. 7 del D.Lgs 196/03 potrà effettuarsi attraverso specifica comunicazione a mezzo posta indirizzata alle medesime Società, o attraverso la casella di posta elettronica dedicata: privacy@rcsdigital.it

Conformemente alla normativa vigente Le chiederemo quindi di esprimere il consenso per i trattamenti di dati barrando la casella "Accetto". Resta inteso che il consenso si riferisce al trattamento dei dati ad eccezione di quelli strettamente necessari per le operazioni ed i servizi da Lei richiesti, al momento della sua adesione in quanto per queste attività il suo consenso non è necessario.

Summary: The UGC is a property of RCS MediaGroup S.p.A., which can use it, and redistribute it to third parties only with the implicit consent of the users. Users have the right to terminate the contract writing to *privacy@rcsdigital.it*.

Rights: Since we have no agreement, this source has been assigned to **DUL3: URLs Only.**

2.3.2.16. Metronews (IT)

URL: http://www.metronews.it/

Terms of Service: No terms of service available online.

Summary: No terms of service can be found on the website

Rights: Since we have no agreement, this source has been assigned to **DUL3: URLs Only**.

2.3.2.17. Il messaggero

URL: http://www.ilmessaggero.it/

Terms of Service: http://www.ilmessaggero.it/privacy/informativa.pdf

Informativa ai sensi del D. Lgs. 196/2003 Ai sensi del D. Lgs. 196/2003 Ced Digital & Servizi S.r.l., società del Gruppo Caltagirone Editore che gestisce le attività internet del Gruppo ed eroga gli abbonamenti alle edizioni digitali, sarà Titolare del trattamento dei dati personali da Lei conferiti, che avverrà nei termini e con le modalità appresso specificati. Se ha sottoscritto un abbonamento con pagamento a mezzo utenza telefonica o carta di credito, La invitiamo a fare riferimento anche all'informativa rilasciata dal suo gestore telefonico o dall'emittente la sua carta di credito. Le ricordiamo che Ced Digital & Servizi S.r.l. non sarà in possesso dei dati della sua carta di credito. Il trattamento dei Suoi dati sarà effettuato manualmente (ad esempio, su supporto cartaceo) e/o attraverso strumenti automatizzati (ad esempio, utilizzando procedure e supporti elettronici) e comunque in conformità alle disposizioni normative vigenti in materia e limitatamente alle finalità specificate nella presente informativa. I Suoi dati personali saranno trattati al fine di erogare il Servizio da Lei prescelto, e compiere le attività ad esso correlate; saranno quindi trattati per l'invio dei prodotti o servizi richiesti, per consentirle di comunicare con la redazione, per la gestione di eventuali reclami e contenziosi, per prevenire frodi, per la tutela e l'eventuale recupero del credito, nonché al fine di realizzare studi statistici aggregati sulla fruizione del Servizio. Il conferimento dei dati è obbligatorio per il conseguimento delle finalità di cui sopra, ed il loro mancato, parziale o inesatto conferimento potrebbe avere come conseguenza l'impossibilità di fornire i servizi o prodotti richiesti. Solo previo consenso informato, e salva la Sua possibilità di opposizione ex art. 7 D. Lgs 196/2003, da esercitarsi nelle forme appresso

specificate, i Suoi dati verranno inoltre trattati per l'invio da parte di Ced Digital & Servizi S.r.l. di comunicazioni di offerte commerciali ed iniziative promozionali o per ricerche di mercato, relative a prodotti editoriali o multimediali del Gruppo Caltagirone, per via telefonica, di posta elettronica o sms/mms, sia a mezzo operatore che in forma automatizzata. La comunicazione dei dati da parte di Ced Digital & Servizi S.r.l. ad aziende terze operanti nei settori informazione ed editoria, telecomunicazioni, e-commerce, mobilità, servizi essenziali, potrà avvenire solo previo consenso informato, sempre salva la suddetta possibilità di opposizione, ed i dati in questo caso verranno trattati dai terzi per le medesime finalità di marketing e con le stesse modalità descritte nel paragrafo precedente. Responsabili del Trattamento potranno essere nominati nelle forme previste dal D. Lgs. 196/2003, per esigenze amministrativo-contabili, di marketing ed analisi di mercato, di natura legale e di tutela del credito. In ogni caso, Lei ha la facoltà di esercitare tutti i diritti previsti dall'articolo 7 del D. Lgs. 196/2003, scrivendo per posta ordinaria a Ced Digital & Servizi S.r.l. con sede legale in via Barberini 28, 00187, Roma oppure inviando un'email all'indirizzo supporto@ceds.it In particolare, l'articolo 7 del D. Lgs. 196/2003 Le garantisce i seguenti diritti:

• ottenere la conferma dell'esistenza o meno di dati personali ed ottenerne la comunicazione in forma intellegibile;

• ottenere l'indicazione dell'origine dei dati personali; delle finalità e delle modalità del trattamento; della logica applicata in caso di trattamento effettuato con l'ausilio di strumenti elettronici; degli estremi identificativi del titolare del trattamento e dei responsabili del trattamento; dei soggetti, o delle categorie dei soggetti, ai quali i dati personali possono essere comunicati o che possono venirne a conoscenza in qualità di responsabili del trattamento o di persone incaricate del trattamento;

• ottenere l'aggiornamento, la rettifica o l'integrazione dei dati personali; la cancellazione, la trasformazione in forma anonima o il blocco dei dati trattati in violazione di legge; l'attestazione che le operazioni indicate in precedenza sono state portate a conoscenza, anche per quanto riguarda il loro contenuto, di coloro ai quali i dati personali sono stati comunicati o diffusi;

• opporsi, in tutto o in parte, al trattamento di dati per motivi legittimi, anche se i dati sono pertinenti allo scopo della raccolta o se sono stati da Lei autorizzati, e quindi revocare il consenso, in tutto o in parte, al trattamento per fini di invio di materiale pubblicitario, di vendita diretta, per il compimento di ricerche di mercato o di comunicazione commerciale, nella forma tradizionale e/o automatizzata.

Summary: The UGC is a property of Ced Digital & Servizi S.r.l., which can redistribute it to third-parties. Users have the right to terminate the contract writing to *supporto@ceds.it*.

Rights: Since we have no agreement, this source has been assigned to **DUL3: URLs Only**.

2.3.2.18. Blog zucconi.repubblica.it

Terms of service: http://quotidiano.repubblica.it/edicola/contratto.jsp?id=38812

I dati, le informazioni, le notizie, i marchi, i loghi, i segni distintivi e, in generale, i contenuti ricevuti dal Fornitore del Servizio in esecuzione del Servizio oggetto del presente Contratto, sono protetti dalle Leggi sulla proprietà intellettuale (Legge 22 aprile 1941, n. 633 e Decreto Legislativo 10 febbraio 2005, n. 30) e, pertanto, sono oggetto di diritti di proprietà intellettuale ed industriale di competenza esclusiva di Elemedia e/o dei suoi danti causa.

L'Utente è pertanto autorizzato ad utilizzare i predetti dati, informazioni, notizie e contenuti, per uso esclusivamente personale nei limiti consentiti dalla Legge e per finalità connesse con la fruizione del Servizio , obbligandosi a non distribuire, disseminare o trasferire con qualsiasi mezzo, anche telematico, ovvero a non pubblicare cedere a terzi a qualsiasi titolo o comunicare al pubblico detti contenuti. Fatto salvo tutto quanto precede, sono vietate le attività di linking, framing, embedding aventi ad oggetto i contenuti, così come ogni altra forma di utilizzazione volta a mettere i contenuti a disposizione del pubblico su internet senza l'espressa autorizzazione scritta del Fornitore.

Summary: Elemedia and/or its licensors are the owners of the intellectual property. The user is authorized to use these data, information, news and content, for personal use only. Activities such as linking, framing, embedding are prohibited without the express written permission of the Supplier.

RIGHTS: We have an internal agreement with the owner of the blog, Vittorio Zucconi, therefore, this data source is assigned to **DUL1: Data under Agreement**, we can edit and share the data.

2.4. IPR scenarios summary table

The following table summarizes the different IPR scenarios found for the data under survey.

Table 2. IPR Scenarios Summary

Source	Written approval?	Use	Edit	Redistribute	UGC
The Guardian, The Independent, The Standard, El Mundo, El Periódico, zucconi.repubblica.it, La Provence, L'express	No (DUL3)	✓	✓	✗	
	Yes (DUL1)	✓	✓	✓	✓
El País, Le Monde, Liberation	No (DUL3)	✓	✓	✗	
	Yes (DUL1/2)	✓	✓	✓	Depends on the written approval
Le Figaro	No (DUL3)	✓	✓	✗	
	Yes (DUL1/2)	✓	✓	✗	Depends on the written approval
Les echos	No (DUL3)	✗	✗	✗	Allowed if anonymized
	Yes (DUL1/2)	✓	✓	✓	
20 minutes	No (DUL1)	✓	✓	✓	✓
	Yes (DUL1)	✓	✓	✓	✓
Metronews.fr	No (DUL3)	✗	✗	✗	✗
	Yes (DUL3)	✓	✓	✗	✓
Corriere della Sera, Il messaggero	No (DUL3)	✓	✓	✗	✗

Source	Written approval?	Use	Edit	Redistribute	UGC
	Yes (DUL1)	✓	✓	✓	✓
microblogging and Social Media Sites	DUL3	✗	✗	✗	✗
Metronews.it	No terms (DUL3)	?	?	?	?

2.5. SENSEI data cases

We have analysed the relevant terms of service of each data source in Section 0 and assigned each data source to their appropriate data usage levels (DULs).

The following table summarises the current state of the DULs across data source owners. Both the content and UGC data from sources with DUL1 can be shared. On the other hand, data with DUL3 can only be referenced and only URLs will be distributed.

Table 3. DULs of SENSEI data sources

Source	Written approval?	Use	Edit	Redistribute	UGC
The Guardian, zucconi.repubblica.it, 20 minutes, micro blogging sites	Yes (DUL1)	✓	✓	✓	✓
microblogging and Social Media sites	DUL3	✗	✗	✗	✗
El País, Le Monde, Liberation, The Independent, The Standard, El mundo, El Periódico, La Provence, L'express, Le Figaro, Les echos, Metronews.fr, Corriere della Sera, Il messaggero, Metronews.it	No (DUL3)	✓	✓	✗	

3. Ethical issues related with collection of human summaries

3.1. Speech

Following what has been done in the Period 1, Teleperformance (TP) has produced, through its Quality Assurance Professional (QAP) team:

1. an “Agent Observation Form”, qualitative monitoring evaluation scheme, for each call center call;
2. a “Call Synopsis” to identify the reason of the every call.

During the listening of the conversations (e.g. from the LUNA or DECODA corpus) the QAP team member takes notes about some specific item, in terms of communications skills of the agent. In order to generate and store the data, each Agent Observation Form is given a numerical code and no personal data is associated or stored.

All TP employees involved in SENSEI project have signed a confidentiality agreement.

3.2. Social Media

In the first period of the project, USFD has gathered feedback from users regarding a set of proposed functionalities (i.e. “use cases”) that SENSEI has identified as of potential utility in the social media setting. Based on the feedback, USFD decided to prioritise the development of summarisation technologies to support Use Case 1: “Town Hall Meeting Summary”.In the second period of the project, USFD evaluated the summarisation technologies developed so far by designing an extrinsic evaluation task of these summarisation technologies (described in D1.3: Report on Intermediate Evaluation). Participants of this task were English native speakers or with excellent English reading skills. They were either professional news producers, such as journalists and editors at local and national newspapers, or members of the public representing interests of news reader and comment providers.

We asked the participants to put themselves in a position of a reader of on-line news and reader comments who has limited time to read a news article and associated reader comments. We then asked them to gain an overview of the comments and to answer questions pertaining to the content of the comments in a series of short, time-limited sessions, by using two different technologies: the Guardian system and the SENSEI system3. After the tasks, participants were invited to take part in a short group discussion regarding their views on the tasks, feedbacks on the systems and ideas of what a future sytem might include. The following data were recorded:

1. Participants’ responses in the pre and post task questionnaires: on paper (which was later scanned and stored electronically).
2. Participants’ spoken responses in a post-task semi-structured dicussion: audio file and paper (researcher’s notes).

[3] This system was developed in the project for presenting reader comment clusters (graphically and textually) and textual summaries, together with linkages to the original comments in context.

Before starting the task, the participants were given an information sheet which provided an explanation about the study and the project, and were asked to sign a consent form to confirm their agreements to participate in this study. They were also informed that they could withdraw at any time should they find the tasks to be too distressing or if they experienced any discomfort during the tasks.

The generated data are used internally to refine and improve the summarisation technologies. All the information that was collected during the course of the study is kept strictly confidential and no identifiable personal information was gathered from the participants for the purpose of these tasks. This study has received ethical approval from the Department of Computer Science ethics committee at the University of Sheffield.

4. Conclusions

We have reported on the management of ethical issues relevant for the social media data collection and distribution. We have identified relevant literatures in this area and evaluated the ethical approaches used in the PANACEA project. We have outlined a Data Usage Level (DUL) categorization scheme for the data sources and its copyright agreements. For each data source, we reported its terms of services and assigned its appropriate DUL, taking into account where an agreement was previously obtained. Whenever the SENSEI consortium will need to distribute data, it will follow the defined data usage level categorization for its data sets. Last but not least, we have implemented the human summary acquisition process for the speech and social media use cases and reported related ethical issues to the tasks.

References

[1] The Misunderstandings of Ownership, Brock Shinen.

http://www.canyoucopyrightatweet.com

[2] The PANACEA project.

http://www.panacea-lr.eu

Appendix 1: Written Approval Letters

Written approval letter model in English

We are a European consortium of several universities, research centers and companies interested in improving technology for understanding human dialogue both spoken and in online conversations. We are working together in a EU funded project called SENSEI[4].

One aspect central to our research is developing technology to better understand and represent conversations arising in newspaper commentaries. To study this we require data from real users interacting with real news papers. This can only be done by downloading commentary threads from online newspapers such as yours.

For the reasons above we request from you permission to:

- Download news and commentary threads from your online news-paper (text only).
- Analyze this content for scientific purposes.
- Publish the results in scientific conferences (this does not entail publishing the data itself, only statistics about the results of the research)

If you agree to this we kindly ask you to email **info@sensei-conversation.eu** with a text similar to this:

> *The newspaper _____ grants the SENSEI research project consortium temporary permission (until 31/10/2016) to download news and commentaries from our online sites (text-only), and indefinite permission to analyze this content for scientific purposes and publish the results (in the form of statistics, not publishing the actual data) in scientific conferences.*

In case that you have agreed to the above, we would further like to ask you to consider helping us further by allowing the distribution of a corpus of news and commentaries from your newspaper for scientific research. This will be part of a much larger corpus of content from European newspapers, social networks, blogs and web pages collected for the project, and would be distributed only for research purposes under written consent. In this case we would ask you to add a text similar to this to the above email:

> *The newspaper further grants the SENSEI research project consortium permission to distribute content from this newspaper's online sites (only text from news and comments) to other research institutions under written consent and only for the purpose of research.*

4 SENSEI: http://www.sensei-conversation.eu (SENSEI FP7-ICT-610916). Consortium: University of Trento (Coordinator), Université d'Aix Marseille, University of Sheffield, University of Essex, Teleperformance and Websays

Written approval letter model in Italian

Vi scrivo per conto di un consorzio formato da Università e aziende Europee nell'ambito del progetto europeo SENSEI[5]. Lo scopo del consorzio è quello di migliorare le tecnologie per la comprensione e l'analisi delle conversazioni, sia nelle conversazioni telefoniche che online. Un aspetto centrale della nostra ricerca è lo sviluppo di una tecnologia per rappresentare e analizzare le conversazioni che avvengono nei commenti agli articoli di quotidiani online. Per farlo utilizziamo dati reali generati da quotidiani come il vostro.

Per queste ragioni Vi chiediamo il permesso di:

- Acquisire *news threads* dal sito online del vostro quotidiano,
- Analizzarne il contenuto per scopi di ricerca scientifica,
- Pubblicare i risultati di questi studi in conferenze scientifiche internazionali (questo non implica la pubblicazione dei dati, ma solo dei risultati della ricerca)

Se siete interessati a consentire questa ricerca o avere più informazioni in merito Vi chiediamo cortesemente di inviare una email all'indirizzo **info@sensei-conversation.eu**

Nel caso siate interessati a consentire questa ricerca Vi chiediamo anche se foste disposti a consentire la redistribuzione dei dati all'interno di un corpus di quotidiani Europei da utilizzarsi unicamente per scopi di ricerca.

Distinti saulti.

Written approval letter model in Spanish

Somos un consorcio europeo de distintas universidades, centros de investigación y empresas interesadas en mejorar las tecnologías para la comprensión del diálogo humano tanto hablado como escrito online. Trabajamos juntos en un proyecto financiado por la EU llamado SENSEI[6].

Un aspecto central para nuestra investigación es el desarrollo de tecnologías para comprender y representar conversaciones que aparecen en los comentarios de los periódicos. Para este estudio es fundamental tener datos reales de usuarios conversando en periódicos reales. Esto solo se puede conseguir descargando las conversaciones que aparecen en periódicos como el suyo.

Por estas razones les pedimos permiso para:

- descargar noticias y sus respectivos comentarios (solo texto)
- analizar este contenido con fines científicos
- Publicar los resultados de la investigación en conferencias (solo publicar resultados, no los datos)

Si están de acuerdo con esto les agradeceríamos que nos enviasen un email a **info@sensei-conversation.eu** con un texto similar al siguiente:

5 SENSEI: http://www.sensei-conversation.eu (SENSEI FP7-ICT-610916). Consortium: University of Trento (Coordinator), Université d'Aix Marseille, University of Sheffield, University of Essex, Teleperformance and Websays

6 SENSEI: http://www.sensei-conversation.eu (SENSEI FP7-ICT-610916). Consortium: University of Trento (Coordinator), Université d'Aix Marseille, University of Sheffield, University of Essex, Teleperformance and Websays

El periódico _____ da permiso al consorcio del proyecto SENSEI permiso temporal (hasta 31/10/2016) para descargar noticias y comentarios de nuestras webs (solo texto), y permiso indefinido para analizar este contenido con fines científicos y publicar sus resultados (solo estadísticas, no los datos) en conferencias.

En caso de que estén de acuerdo con el apartado anterior, nos gustaría pedirles que consideren ayudarnos aún más dándonos permiso para publicar y distribuir los datos con las noticias y los comentarios con fines científicos. Estos datos formarían parte de un corpus más grande con contenido de periódicos europeos, redes sociales, blogs, etc. que sería distribuido solo bajo consentimiento escrito y con fines científicos. En este caso, os pediríamos que añadieseis al email anterior el siguiente fragmento:

El periódico da además permiso al consorcio del proyecto SENSEI para distribuir contenido (solo texto de noticias y comentarios) a otras instituciones dedicadas a la investigación con permiso escrito y solo con fines científicos.

Written approval letter model in French

Objet : Autorisation d'utilisation des données à des fins de recherche

Je vous écris, en tant que membre du projet de recherche européen SENSEI, pour vous demander l'autorisation d'utiliser des données publiées sur votre site web à des fins de recherche.

Nous sommes un consortium européen d'universités, centres de recherches et entreprises, intéressés par l'amélioration des technologies pour la compréhen-sion automatique des dialogues humains, en particulier les interventions d'inter-nautes. Nous travaillons ensemble dans le cadre d'un projet de recherche euro-péen, le projet SENSEI.

Un aspect central de nos recherches est le développement de technologies pour mieux comprendre et représenter les conversations qui prennent place dans les commentaires à des articles d'actualité sur le web. Un intérêt de ces recherches est qu'elles étudient des interactions réelles entre de vrais utilisateurs. Ce genre d'étude ne peut être effectuée qu'en téléchargeant et traitant les fils de discus-sions depuis les sites web de journaux comme le votre.

Les bénéfices attendu des résultats du projet SENSEI sont une meilleure compréhension du comportement des internautes sur les sites de journaux et un ensemble d'outils utilisables par les journalistes pour fouiller de grandes quantités de contributions utilisateurs : sur quels sujets portent les opinions exprimés par les internautes ? De nouvelles informations émergent-t-elles des commentaires ? Qu'attendent ces internautes des journalistes ?

Pour ces raisons, nous vous demandons la permission de :

- télécharger le contenu des articles et commentaires depuis votre site web (uniquement le contenu textuel) ;
- analyser ce contenu à des fins de recherche scientifique ;
- publier les résultats de ces recherches dans des conférences et revues scientifiques (ceci n'implique pas la publication des données elles-mêmes, seulement des statistiques à propos des résultats scientifiques).

Si vous acquiescez à cette demande, nous vous demandons de bien vouloir adresser par email à info@sensei-conversation.eu un texte similaire à ce dernier :

Le journal _____ autorise les membres du projet de recherche SENSEI de manière temporaire (jusqu'au 31 octobre 2016, fin du projet) à télécharger les articles et commentaires à partir du site web _____ (texte uniquement), et une permission sans limite de durée d'analyser ce contenu à des fins de recherche scientifique et à publier les résultats de ces recherches dans des conférences et revues scientifiques.

Dans le cas où vous accepteriez les conditions précédentes, nous vous demandons de considérer de nous aider un peu plus et nous permettre de distribuer un corpus d'articles et de commentaires associés construit depuis votre site web, toujours à des fins de recherche. Ce contenu fera partie d'un corpus beaucoup plus gros for-mé de données collectées sur des sites de journaux européens, des réseaux sociaux, des blogs et des pages web, dans le cadre du projet SENSEI. Ce corpus serait distribué sur accord écrit et à des fins de recherche uniquement. Si vous acceptez, merci d'adapter le texte suivant et l'ajouter à votre email.

Le journal _____ autorise les membres du consortium SENSEI à distribuer le contenu téléchargé (texte des articles et commentaires) à d'autres institutions de recherche après permission écrite et uniquement à des buts de recherche scientifique.

SENSEI: http://www.sensei-conversation.eu (SENSEI FP7-ICT-610916). Consortium : Université de Trento (coordinateur du projet), Université d'Aix-Marseille, Université de Sheffield, Université d'Essex, Teleperformance, Websays.