Test material including ground truth v3

Deliverable D4.2.3

TOSCA-MP identifier:	TOSCAMP-D4.2.3-VRT-v1.00.docx
Deliverable number:	D4.2.3
Author(s) and company:	Mike Matton (VRT)
Internal reviewers:	EBU
Work package / task:	WP4
Document status:	Final
Confidentiality:	Public

Version	Date	Reason of change
0.1	2013-12-19	Document created, starting from D4.2.2
0.2	2014-01-06	Added details of content
0.9	2014-01-14	Added details on concept detection ground truth
0.91	2014-02-04	Added final content and updated statistics
1.00	2014-02-13	Final version after internal review

Acknowledgement: The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 287532.

Table of Contents

List of Figures

List of Tables

1 Executive Summary

This document accompanies the release of version 3 of the mammie platform. Mammie is a platform which is used to distribute test media material in the TOSCA-MP consortium. It is available through a website located at http://tosca-mp.vrt.be/toscamp/.

The mammie platform is a web interface to a lightweight media asset management system. It contains a few basic MAM functions, such as access control, ingest, update and download of media assets and the corresponding metadata streams. The metadata is searchable using a basic search engine.

This deliverable contains all material that was already present in version 2. In this version, some more content was added and ground truth annotations have been provided for some of the content by different partners.

The mammie platform currently contains 5,429 videos which have been provided by the different content providers. This provides a total duration of nearly 1,900 hours of high resolution broadcast video currently available on the platform. The contributions of the different partners namely IRT, RAI and VRT are distributed as follows:

Partner	Number of media items	Total duration
IRT	2923	597 hours
RAI	1541	845 hours
VRT	965	438 hours

Table 1 - Partners contribution overview

The platform is accompanied by a material contract, which each TOSCA-MP partner has to sign before requesting access to the platform.

Furthermore, ground truth material has been made available. For a selected portion of the content, speech transcriptions, text translations, visual concepts, genre classification, quality control and shot boundary segmentation has been annotated and made available to the consortium. We refer to section 4.3 for detailed statistics.

2 Introduction

2.1 Purpose of this Document

Document that accompanies the third release of test media material on the mammie platform.

2.2 Scope of this Document

This deliverable covers the third release of test media material to be shared with the consortium.

2.3 Status of this Document

This is the final version of D4.2.3

2.4 Related Documents

Material contract (see document store).

3 The mammie platform

This deliverable consists of the mammie software platform. Mammie is a lightweight media asset management system which is used to distribute content between the partners in the TOSCA-MP consortium.

3.1 Access

The platform is accessible through a web interface. The url for the web interface is
http://tosca-mp.lab.vrt.be/toscamp

Requests for access to the platform can be directed to VRT, Mike Matton <mike.matton@vrt.be>

3.2 Authentication

Upon opening the web interface, the user is redirected to the authentication page. All content is protected by a username/password verification. The partners in the TOSCA-MP consortium can request for access to the platform after signing a material contract that has been specifically created for the TOSCA-MP partners. A link to this material contract can be found on the bottom of the mammie web interface.

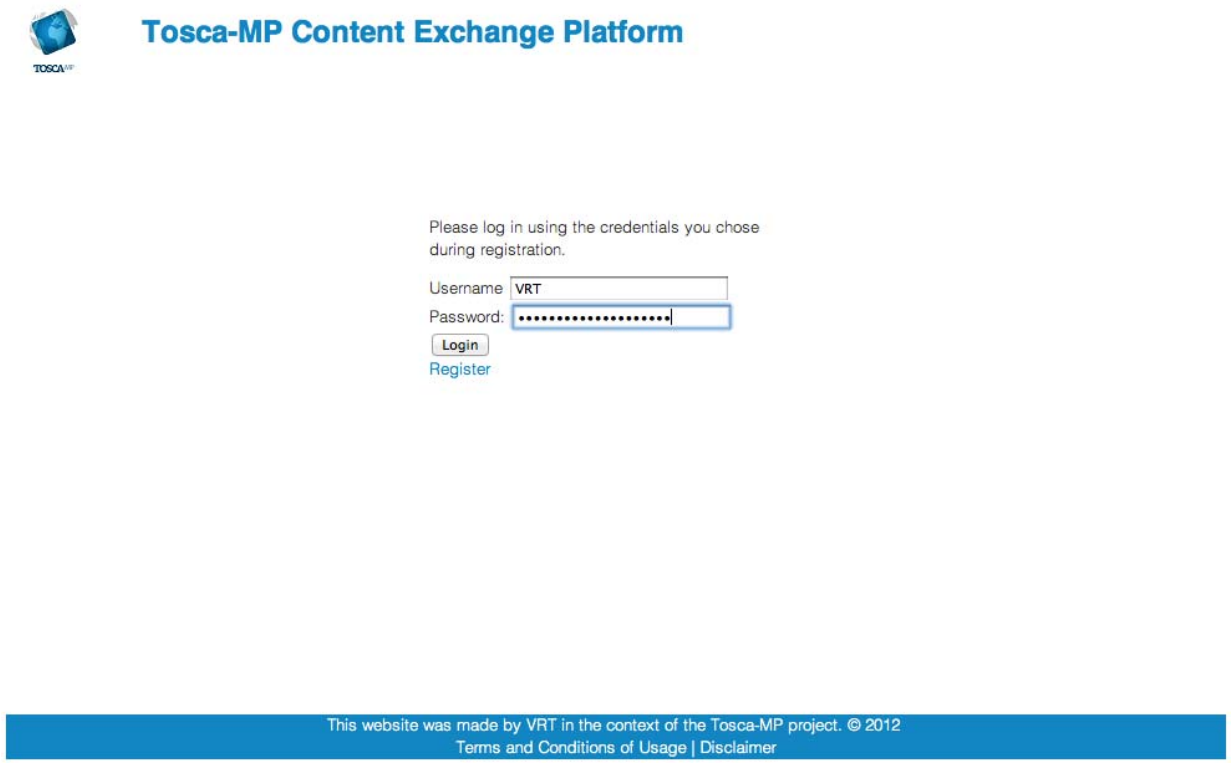

Figure 1 - mammie login screen

3.3 Usage

3.3.1 The overview page

After logging in, the user is presented with an interface displaying all available content on the platform.

On the top right, the different contributors are listed (namely: RAI, IRT and VRT). Next to the name, the number of available videos is shown. A user can click on a particular content provider to select only the content available from that content provider.

On the top of the page is a search field. Using this search field, the user can search through the available metadata for the video. It is not intended to be a fully fledged search functionality, but only as a way to quickly select a particular set of videos.

The main window shows an overview of the selected videos, paginated in sets of 10 results per page. Upon a click on one particular result, the detailed view of that item is opened.

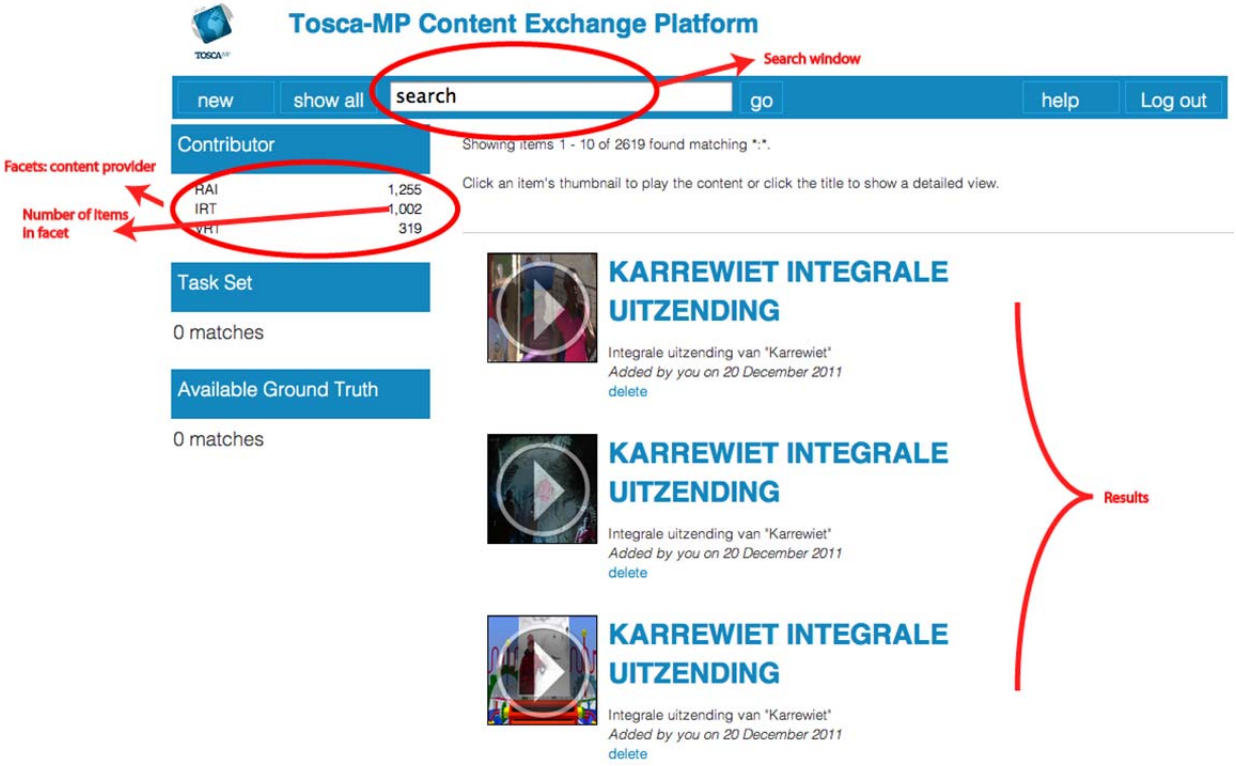

Figure 2 - mammie overview page

3.3.2 The detailed view

The detailed view of an item shows a media player that can by used to play a browse version of the content. Around this media player, the available metadata is visualised. The metadata is stored in Dublin Core format.

Below this, a list of the available content files is shown. Currently, the available media files are the high-resolution video, and two low-resolution browse versions, encoded in h264 and ogv. These files can be downloaded with a click on the corresponding content file.

On the bottom of the page, the available metadata streams are visualised. For every item, the Dublin core and some technical metadata can be downloaded in XML format. Moreover, for some files from VRT, the subtitles of the video are also available in XML format.

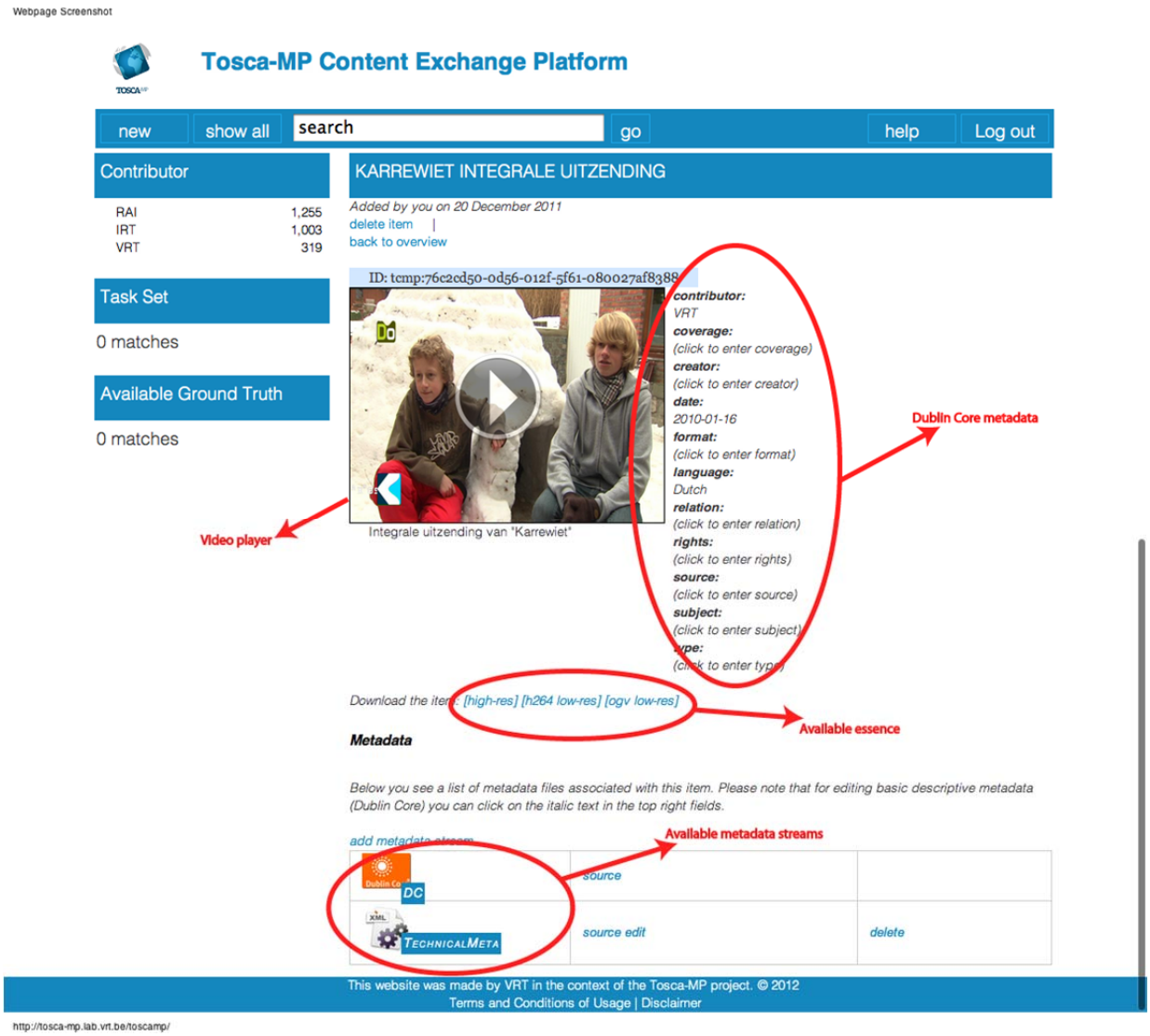

Figure 3 - mammie detailed view

4 Content available in mammie platform

4.1 Introduction

In this section, an overview of the material in the first version of the platform is presented. Section 4.2 contains an overview of the items contributed per partner. In total there is almost 800 hours of broadcast video currently available on the platform.

4.2 Available content

In total, there currently are 5,434 videos on the platform in this first iteration of the deliverable. Below is a summary of the videos provided by the different partners.

4.2.1 IRT

IRT has provided podcasts. There currently are 1238 podcasts on the platform.

Content type	Number of items	Total duration
Angela_Merkel_-_Die_Kanzlerin_direkt	43	2h52’40”
ARD-Beiträge_Grubenunglück_Chile	21	1h20’35”
Bericht_aus_Berlin/2009	12	3h39’54”
Bericht_aus_Berlin/2011	8	2h26’24”
BR_Abendschau_-_Politik_&_Wirtschaft	435	22h24’31”
Europa_Aktuell_ Das_Magazin_aus_Brüssel	128	10h34’17”
Journal_Interview	115	17h14’21”
N24_Kompakt	476	40h07’34”
Nachtmagazin/2008-2009	86	28h13’37”
Nachtmagazin/2011	36	12h04’50”
n-tv_Nachrichten	180	22h26’04”
PHOENIX_Im_Dialog	29	14h17’41”
PHOENIX_Unter_den_Linden	33	24h53’53”
Politik_direkt_-_Das_Politikmagazin	150	12h54’34”
Quarks_&_Co	64	44h14’07”
Tagesschau/2008-2009	164	41h06’36”
Tagesschau/2011	81	20h23’29”
Tagesthemen/2008-2009	169	60h14’03”
Tagesthemen/2011	80	28h46’52”
TTT_-_Titel_Thesen_Temperamente	36	0h52’59”
Wochenspiegel/2008-2009	15	7h31’31”
Wochenspiegel/2011	4	1h59’12”
ZDF_-_heute	48	14h15’49”
ZDF_-_heute-journal/2008-2009	249	99h45’45”
ZDF_-_heute-journal/2011	48	18h29’31”

ZDF_-_logo_Deine_Nachrichten	186	29h32’03”
Wedding data set	5	1h46’34
SCAIE podcasts	41	17h43’00”

Table 2 - IRT contributed content

4.2.2 RAI

Currently there are only recent broadcast news shows available on the mammie platform. This is the summary:

Content type	Number of items	Total duration
Recent broadcast news	1222	620h34’54”
Archive broadcast news	30	55h50’15”
Talkshow 90esimo Minuto	20	5h37’28”
Talkshow AnnoZero	48	25h11’18”
Talkshow Ballaro	49	25h46’13”
Talkshow Che Tempo Che Fa	84	25h47’20”
Talkshow Porta Porta	40	26h17’40”
Other talkshows	19	27h32’53”
Sports content	8	8h00’00”
Various	21	22h29’51”

Table 3 - RAI contributed content

For all newscasts, RAI has also provided metadata obtained with their ANTS system for automated metadata extraction.

4.2.3 VRT

VRT has provided broadcast news shows from 2010 for the project. This is a summary of the available content:

Content type	Number of items	Total duration
Broadcast news from 2010	694	324h49’33”
Broadcast news Kate & William	27	11h47’44”
Newsrushes	153	1h32’04”
Talkshow De Laatste Show	10	8h08’11”
Talkshow De Zevende Dag	19	37h32’47”
Talkshow Volt	26	24h32’57”
Magazine Vlaanderen Vakantieland	27	20h01’15”
Qc material	8	4h40’25””
Sports	1	3h10’15

Table 4 - VRT contributed content

Where available, VRT has also provided subtitles. These have been added for almost all available newscasts.

4.3 Available ground truth annotations

4.3.1 Speech transcriptions

The speech transcriptions were generated through crowd sourcing. For more information about the procedure, we refer to TOSCA-MP deliverable D2.2. The following table lists an overview of the generated transcriptions.

	News broadcast	**Talk-show**	**Total**
German	4h03m (13)	5h02m (8)	9h05m (21)
Italian	3h54m (8)	7h21m (5)	11h16m (13)
English	5h07m (11)	0h0m (0)	5h07m (11)
Dutch	5h38m (13)	6h07m (5)	11h45m (18)

4.3.2 Machine translation

	Translation direction			
	NL → EN	**EN → IT**	**DE → EN**	**DE → IT**
# of words	20,744	20,069	21,179	21,179

4.3.3 Concept detection

The following concepts were annotated for 26 selected videos: anchor, female, indoor, interview, male, outdoor, person, studio. This resulted in 6106 annotated key frames.

Furthermore, for 45 videos of the wedding data set, the following concepts and actions have been annotated: crowd, vehicle passing by, interview, building, parked cars, accident, fire.

4.3.4 Genre classification

The genre was annotated for 1437 items in de content set, divided into three basic genres: news (1188 items), sports (8 items) and talk shows (241 items).

4.3.5 Quality control

For a few videos, ground truth aboud sharpness is available. This ground truth has been generated as part of the pilot field trials, which is documented in TOSCA-MP deliverable D6.4.1.

4.3.6 Shot boundary detection

Ground truth for shot boundaries has been created by pooling results from automatic analysis methods from RAI, HHI and JRS. All analysis results have been represented using MPEG-7 AVDP, and the tools developed in WP4 have been used to determine differences between the documents, and classify to different types of errors. Deviations of shot boundaries in the range of 5 frames have been accepted. Other deviations are marked in the merged results, so that this information can be considered when using the data for evaluations, or perform corrections on the data. The shot boundary information has been created for 1165 files of the data set on MAMMIE, resulting in about 450K shot boundary annotations.

5 API interface

The mammie system also has an REST API-interface available for integration purposes. For a description of this API, we refer to public TOSCA-MP deliverable D3.2.

6 Glossary

Terms used within the TOSCA-MP project, sorted alphabetically.

see D1.2 project handbook, version 3 or later

Partner Acronyms

DTO	Technicolor, DE
EBU	European Broadcasting Union, CH
FBK	Fondazione Bruno Kessler, FBK
HHI	Heinrich Hertz Institut, Fraunhofer Gesellschaft zur Förderung der Angewandten Forschung e.V., DE
IRT	Institut für Rundfunktechnik GmbH, DE
K.U.Leuven	Katholieke Universiteit Leuven, BE
JRS	JOANNEUM RESEARCH Forschungsgesellschaft mbH, AT
PLY	Playence KG, AT
RAI	Radiotelevisione Italiana S.p.a., IT
VRT	De Vlaamse Radio en Televisieomroeporganisatie NV, BE

Acknowlegdement: The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 287532.

