

Alexander Arturo Mera Caraballo

**Publishing Annotated Multimedia  
Deep Web Data**

**DISSERTAÇÃO DE MESTRADO**

**DEPARTAMENTO DE INFORMÁTICA  
PROGRAMA DE PÓS-GRADUAÇÃO  
EM INFORMÁTICA**

Rio de Janeiro

April 2012



**Alexander Arturo Mera Caraballo**

## **Publishing Annotated Multimedia Deep Web Data**

### **Dissertação de Mestrado**

Dissertation presented to the Programa de Pós-graduação em Informática of the Departamento de Informática, PUC-Rio as partial fulfillment of the requirements for the degree of Mestre em Informática.

Advisor: Prof. Marco Antonio Casanova

Rio de Janeiro, April 2012.



**Alexander Arturo Mera Caraballo**

## **Publishing Annotated Multimedia Deep Web Data**

Dissertation presented to the Programa de Pós-graduação em Informática of the Departamento de Informática, PUC-Rio as partial fulfillment of the requirements for the degree of Mestre em Informática

**Prof. Marco Antonio Casanova**

Orientador

Departamento de Informática -- PUC-Rio

**Prof. Antonio Luz Furtado**

Departamento de Informática -- PUC-Rio

**Prof. Luiz André Portes Paes Leme**

Instituto de Computação -- UFF

**Prof. José Eugenio Leal**

Coordenador (a) Setorial do Centro

Técnico Científico - PUC-Rio

Rio de Janeiro, April 03, 2012.

**Alexander Arturo Mera Carballo**

Graduated in Systems Engineering from Universidad de Nariño (UDENAR), Pasto – Colombia in 2009.

Bibliographic data

Mera Carballo, Alexander Arturo

Publishing annotated multimedia Deep Web data/  
Alexander Arturo Mera Carballo; Advisor: Marco Antonio Casanova – 2012.

67f.: il. (Color) ; 30 cm

Dissertação de (mestrado) – Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Informática, 2012.

Inclui bibliografia

1. Informática – Teses. 2. Recuperação de conteúdo falado. 3. Recuperação de dados multimídia. 4. Objetos de aprendizagem. 5. Navegação de áudio. 6. Recuperação de informação multilíngue. I. Casanova, Marco Antonio. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. III. Título.

CDD: 004

## Acknowledgments

I wish to thank, first and foremost, my advisor Prof. Marco Antonio Casanova, whose encouragement, guidance and everyday kindness made possible the realization of this work.

I also would like to make a special reference to MSc. Bernardo Pereira Nunes for his friendship, valuable insights and recommendations.

I want to thank to my beloved family, for their patience and support.

I remain indebted to many colleagues and professors at PUC-Rio for providing me the means to learn and understand.

To CAPES, for the financial support, without which this work would not have been possible

## Abstract

Mera Caraballo, Alexander Arturo; Casanova, Marco Antonio. **Publishing Annotated Multimedia Deep Web Data**. Rio de Janeiro, 2012. 67p. MSc Dissertation - Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

In recent years, we witnessed a huge growth of multimedia data on the Web. New lower-cost technologies and greater bandwidth allowed the Web to evolve into a multimedia format. However, the lack of tools that can make multimedia format easily accessible on the Web led us to a non-searchable and non-indexable data of the Web, also known as Deep Web. In line with these observations, this dissertation addresses the problem of how to publish audio and video content on the Web. We present a tool and a novel approach that facilitates the indexing and retrieval of the objects with the help of traditional search engines. The tool automatically generates static Web pages that describe the content of the objects and organize this content to facilitate locating segments of the audio or video which correspond to the descriptions. The static Web pages can be translated to others languages to reach other user populations. An annotation process is also performed to embed machine-readable data into the Web pages. The dissertation also presents an in-depth experiment, publishing learning objects based on audio and video, to assess the efficacy of the technique.

## Keywords

Spoken content retrieval; Multimedia retrieval; Learning objects; Audio browsing; Cross-Language Information Retrieval.

## Sumário

Mera Caraballo, Alexander Arturo; Casanova, Marco Antonio. **Publicando anotações de dados multimídia advinda da Deep Web**. Rio de Janeiro, 2012. 67p. Dissertação de Mestrado - Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Nos últimos anos, temos assistido um enorme crescimento de dados multimídia na Web. Novas tecnologias de menor custo e maior largura de banda têm permitido que a Web evolua para um formato multimídia. No entanto, a falta de ferramentas que podem tornar o formato multimídia disponível na Web nos levou a um conjunto de dados não-pesquisável e não indexável da Web, também conhecido como Deep Web. Desta forma, esta dissertação aborda o problema de como publicar conteúdo de áudio e vídeo na Web. Apresentamos uma ferramenta e uma nova abordagem que facilita a indexação e recuperação dos objetos com a ajuda das máquinas de busca tradicionais. A ferramenta gera automaticamente páginas Web estáticas que descrevem o conteúdo dos objetos e organizar esse conteúdo para facilitar a localização de segmentos do áudio ou vídeo que correspondem às descrições. As páginas Web estáticas podem ser traduzidos para outras línguas para atingir outras populações de usuários. Um processo de anotação também é realizado para incorporar dados legíveis pelas máquinas nas páginas Web. A dissertação também apresenta um experimento completo, publicando objetos de aprendizagem baseados em áudio e vídeo para avaliar a eficácia da abordagem.

## Palavras-chave

Recuperação de conteúdo falado; Recuperação de dados multimídia; Objetos de Aprendizagem; Navegação de Áudio; Recuperação de Informação Multilíngue.

# Contents

1	Introduction	12
2	Background and Related Work	14
2.1	A Classification for Multimedia Metadata	14
2.2	Video Annotation	15
2.2.1	Text-Based Approaches	15
2.2.2	Audio-Based Approaches	20
2.2.3	Visual-Based Approaches	23
2.3	Image Annotation	25
2.3.1	Text-Base Image Retrieval	25
2.3.2	Content-Base Image Retrieval	27
2.4	RDF-a	29
2.5	Related Work	30
3	Publishing Technique	32
3.1	Overview of the Publishing Technique	32
3.2	Stage 1 - Automatic speech recognition	34
3.3	Stage 2 - Translation	34
3.4	Stage 3 - XHTML-ize	36
3.5	Stage 4 - Annotation	37
4	Publishing Tool	39
4.1	Overview of the Publishing Tool	39
4.2	Details of the Publication Tool	41
5	Experiments	53
5.1	Experimental Setup	53
5.2	Data Analysis	54
5.2.1	An analysis of total number of hits	54
5.2.2	An analysis of the number of hits by language	54
5.2.3	An analysis of the number of hits by asset	57
5.2.4	Regional analysis	57



6	Conclusions and Future Work	59
7	Bibliography	60
A	SubRip text file for the user Publication example	63

## List of Figures

Figure 1(a) (b) Scenes text images.	16
Figure 2 Shows captions directly on the background.	16
Figure 3 (e) (f) Multi-color document images.	16
Figure 4 Architecture of a TIE system [9].	18
Figure 5 Example speech recognition lattice	20
Figure 6 Decomposition of an audio signal into frames and clips [12].	21
Figure 7 Scene classification. Low level.	28
Figure 8 Publishing technique architecture.	33
Figure 9 Flowchart of processing for performing publication in accordance with the implementation.	40
Figure 10 SubRip file text.	42
Figure 11 Publishing tool interface.	43
Figure 12 Processing a A/V asset.	44
Figure 14 Request to translate English caption to Spanish.	45
Figure 15 Request to translate English caption to Portuguese.	45
Figure 16 Extraction of content-independet Metadata.	46
Figure 17 XHTML-ize process on the metadata.	46
Figure 18 XHML-ize process on the transcription.	47
Figure 19 Dublin Core enrichment.	48
Figure 20 Diagram representation of the annotation in Figure 19	48
Figure 21 Request to the SpotLight service endpoint.	49
Figure 22 Example of an enriched Web document.	50
Figure 23. Graph representation of the annotated content in Figure 22.	50
Figure 24 [N3] sintax of the Graph in .	51
Figure 25 HTML representation of the annotation stage.	51
Figure 26 Spoken Content page.	52
Figure 27 Hit stats. First stage 1–5. Second stage 5–8.	54
Figure 28 Hits percentage of translated static Web pages generated by our publishing technique. 37% of the hits were obtained by the Web pages of the assets in their native language and 67% of the hits were obtained by the translated Web pages.	55
Figure 29 Countries that have interacted with the content: Japan, Brazil and Portugal.	56

Figure 30 Countries that have interacted with the content: Sweden, Japan, Brazil, Spain, Peru, USA and Portugal. 56

Figure 31 The number of hits boosted in different orders of magnitude for each continent. 58

## List of Tables

Table 1 Translated transcription into Spanish and Portuguese.	45
Table 2 Top 10 LO's by number of hits.	56
Table 3 Last 10 least hit Learning Objects.	57