1
Introduction

Deep Web refers to World Wide Web content that is not directly indexed by search engines. Studies showed that Deep Web content is particularly important. However, obtaining such content is challenging and has been acknowledged as a significant gap in the coverage of search engines [1].

The Web is rapidly evolving into a multimedia format. The number of multimedia data has increased dramatically since the technology became easier to use at low cost. The accessibility and ability to search within this large archive are limited and difficult [2]. This problem is mostly due to the fact that the content of multimedia data is often not available in machine-readable form and is described with the help of metadata. Therefore, search engines, robots or agents can process multimedia data only by accessing metadata [3].

In this dissertation, we focus on what we call *spoken content* on the Web, that is, audio and video (with audio) objects. Unlike text-based content, spoken content cannot be easily indexed. Furthermore, any reasonable publishing technique for spoken content must help users locate the time-aligned segments within the content that meets their needs.

We describe a technique and implement a tool to create textual descriptions for spoken content that transforms the spoken search problem into a traditional text search problem. In addition, the tool processes spoken content in such a way that text-based search engines may help users locate time-aligned segments of the content. Furthermore, the tool annotates entities [4], present in the transcriptions, from a specific set of ontologies in DBPedia [5] using RDFa. To demonstrate the effectiveness of the tool, we describe an experiment with more than a thousand minutes of spoken contents, divided into 99 video objects.

The rest of this dissertation is organized as follows. In chapter 2, we review the background concepts of the annotation process in multimedia data and discuss the related work. In chapter 3, we introduce the publishing technique. In

chapter 4, we describe the implementation of the publishing tool. Finally, in chapter 5, we discuss the conclusions and the challenges to be met in the future.