Chapter 11
NLP and Digital Library Management

Lyne Da Sylva
University of Montreal, Canada

ABSTRACT

The field of study of Natural Language Processing (NLP) has developed over the past 50 years or so, producing an array of now mature technology, such as automatic morphological analysis, word sense disambiguation, parsing, anaphora resolution, natural language generation, named entity recognition, etc. The proliferation of large digital collections (evolving into Digital Libraries) and the emerging economic value of information demand efficient solutions for managing the information which is available, but which is not always easy to find. This chapter presents the requirements for handling documents in digital libraries and explains how existing NLP technology can be used to facilitate the task of document management.

INTRODUCTION

The field of study of Natural Language Processing (NLP) has developed and ripened in the past 50 years or so, from the first machine translation and information retrieval applications to the present. These two areas of research have been far-reaching and pervasive. In the process of resolving issues of understanding natural language, for both translation and retrieval, many sub-areas of NLP have emerged: automatic morphological analysis, word sense disambiguation, parsing, anaphora resolution, natural language generation, named entity recognition, etc.

In today's research in NLP, attention has shifted from machine translation over to different versions of Information Retrieval (IR) applications. The increasing availability of large collections of digital documents has spurred interest in devising useful technology to handle these. Specifically, the notion of “digital libraries” (Adams, 1995; Fox, et al., 1995; Arms, 2000) has emerged, with specific architecture and functionality. This is an area where many mature NLP applications can be brought into play. It is an area mostly associated with IR, which has traditionally used little NLP and yet produced efficient tools; methods needed to include more sophisticated, NLP-based approaches were, up to recently, beyond the reach of IR systems. But digital libraries are much more than simply IR.
This chapter has the following three objectives: (1) to describe the issues relating to the task of managing a digital library; (2) to explore various NLP applications which can be applied to the task; (3) to identify new research problems related to these issues.

BACKGROUND: DIGITAL LIBRARIES, DOCUMENT MANAGEMENT, AND NLP

Digital Libraries

Digital collections existed long before the advent of the Web and the coinage of the term “digital library.” NetLib (http://www.netlib.org/), created in 1985, contains a collection of freely available software, documents, and databases of interest to the numerical, scientific computing, and other communities. The Perseus project (http://www.perseus.tufts.edu/hopper/) was created in 1985 to host a collection of resources on Ancient Greece: documents, images of artefacts, maps and the like, all linked together to allow a better understanding of Ancient Greek texts. Cornell University’s e-prints archive (http://arxiv.org/), formerly the Los Alamos E-print Archive, dates from 1991. It contains prepublications in the field of physics and related disciplines. These are but a few examples among many. They were, however, isolated initiatives, suffering from minimal interfaces providing access to resources over less than efficient networks. Improvements in interface design and in network configurations, the advent of the WWW and increasing publication of materials on the Web led naturally to the creation of communities of users wishing to share and publish resources – and of technology to support it.

From a computer science perspective, digital libraries are an extension of network technologies, databases and search engines. From an information science viewpoint however, digital libraries are institutions and not machines; they are a logical extension of traditional libraries, whose mission is to acquire, organise and disseminate information. They also mean other things to other groups: a new outlet for content providers, publishers, museums, and commercial vendors; a democratization tool for governments; a new service channel for educators. In addition, from the viewpoint of NLP they represent a new opportunity, a new area of application in which to deploy existing technology, perfect it and invent more.

Definitions for what constitutes a digital library are many, and reflect the fact that work on digital libraries stems from a number of different fields, including computer science and information science of course. Relevant literature on new research is to be found in topical conference proceedings: the European Conference on Digital Libraries (ECDL), the Joint (ACM-IEEE) Conference on Digital Libraries, the International Conference of Asian Digital Libraries, the International Conference on Digital Libraries and the new Theory and Practice of Digital Libraries (formerly ECDL). It also is present in library association conferences and pre-existing conferences of information scientists, publishers, abstracting and indexing services, and online database providers (Bearman, 2008).

An early definition, still cited today, comes from Borgman (2000, p. 42), in which a digital library is as follows:

... A set of electronic resources and associated technical capabilities for creating, searching, and using information. In this sense they are an extension and enhancement of information storage and retrieval systems that manipulate digital data in any medium (text, images, sounds; static or dynamic images) and exist in distributed networks. The content of digital libraries includes data, metadata that describe various aspects of the data (e.g. representation, creator, owner, reproduction rights), and metadata that consist of links or relationships to other data or metadata, whether internal or external to the digital library.