Ethical Issues in Web Archive Creation and Usage – Towards a Research Agenda

Andreas Rauber
Vienna University of Technology
1040 Vienna, Austria
http://www.ifs.tuwien.ac.at/~andi
rauber@ifs.tuwien.ac.at

Max Kaiser
Austrian National Library
Josefsplatz 1
1010 Vienna, Austria
max.kaiser@onb.ac.at

Bernhard Wachter
Vienna University of Technology
1040 Vienna, Austria
bernhard.wachter@gmail.com

ABSTRACT

While Web archiving initiatives rescue a wealth of information on the Web from being permanently lost, the massive collection of Web data poses not only fascinating possibilities for accessing a vast amount of information, as well as an invaluable resource for scientist wanting to understand the technological and sociological development of the Web and society at large. It also constitutes a new type of information on its own, posing numerous ethical challenges, specifically given the powerful techniques for analyzing and exploring the masses of accumulated information that we will have available in the near future.

Being aware of this issue, most Web archives currently strictly limit access to their holdings, or provide means to allow people having their content excluded from holdings to avoid the subsequent challenges, at the same time drastically limiting their value and usefulness.

This paper discusses some of the key concerns that may be validly raised in opposition to Web archiving initiatives, and points out directions requiring further research to pro-actively address these concerns, with a focus on IT-related aspects. We further report on exemplary studies trying to automatically identify personal text segments in Web pages as an initial step in addressing one facet of the challenges identified.

Keywords

Web Archiving, Access, Ethical Issues, Information Retrieval, Research Challenges

1. INTRODUCTION

The Web archives that are being created by many national libraries and several national or specialist archives and libraries across the world constitute an invaluable source of information. They serve as reference basis for Web pages and documents that are no longer on-line, but also as a rich body of information as a whole, documenting the evolution of the information society. However, access to these archives is currently severely limited and restricted. On the one hand this is still due to the lack of tools supporting flexible means of interaction with the large bodies of data. Several initiatives such as the Nordic Web Archive Access Tools, WERA, the Wayback Machine [11], and the recent FP7 project LIWA are working on solutions to overcome this challenge. However, there are still many aspects in providing access to huge Web archives which need to be researched.

Additionally, evaluations of existing Web archives show that still they can be opened only to a very limited degree and to a highly selected fraction of users (mostly to researchers, which in turn are able to demonstrate that highly valuable information can be gleaned form these archives – as do usage numbers for those archives that are publicly available, such as the Internet Archive) due to ethical and legal reasons. Web archiving initiatives have realized, that the body of information represented by their archives constitutes not only a simple body of factual information, but that they represent a novel type of collections that may also be utilized and abused in ethically and legally questionable ways. Not only due to copyright reasons, but also due to privacy and data protection considerations, these valuable holdings remain sealed off from public access, or - at best - constitute isolated islands of national content, breaking at the national boundaries, and thus completely losing the potential as well as the characteristics of the very medium that they are based upon, i.e. a highly interconnected network of information and, ultimately, society. This is predominantly due to the fact that access regulations in the various countries are different, effectively prohibiting networked access to the content they are holding across national boundaries. An overview of the various regulations governing Web archiving is provided in [5]. Global initiatives, on the other hand, inevitably are limited to a much shallower or less frequent coverage of Web content when collecting information on a global scale.

Ethical issues in computing, and specifically with respect to activities in the Internet, have been receiving considerable attention. Recognizing the need to ensure a balance between the protection of human subjects as well as the promotion of sound research, a specific workshop was organized in 1999 to analyze the ethical and legal aspects of human subjects research on the internet [3]. When it comes to access and search within large volumes of data, privacy protecting data mining techniques are being developed [12]. Equally broad interest is devoted to the responsibilities of search engines, both from a legal as well as ethical perspective. Many of the problems raised in this field are also applicable to the domain of Web Archiving. These address issues such as copyright infringement, or potential impact of ranking on information provision, as well as issues related to providing access to wrong or outdated information. To counter the latter, concepts such as the right to provide a “reply” to the information returned by a search engine, have been proposed [6],