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],

which may be an option to resolve critical issues, but probably do not constitute a scalable solution.

Still, we find most research efforts in the context of Web archiving being focused on improving data collection and management as well as, specifically, more powerful means of access to unlock their value – while exactly this access is being limited due to ethical or legal concerns. One notable exception to this is a presentation by Foot et al. analyzing the ethics of Web archiving [2]. They identify numerous aspects on all levels, ranging from the creation of a Web archive that may constitute a new type of data collection, to the provision of access and potential abuses.

While the ultimate solution to the problem of what kind of access will be permissible will have to be a legal one, it is important to understand the detailed characteristics of the possibilities of "unethical" exploitation of a Web archive's holdings, or simply types of usage that some people may feel uncomfortable with concerning the content that they have made available on the Web sometime in the past. It requires a profound understanding of the semantic and cognitive aspects and values of the information aggregated over time, as opposed to the individual pages it is based upon and that are currently searchable via conventional Web search engines.

It will also require the development of technical means to counter these challenges. This includes researching the potential of new techniques of large-scale information retrieval including its semantic capabilities and, specifically, the drastically different level of semantic information that can be gleaned from the unprecedented collection of information that is available in Web archive holdings. This paper thus aims at providing a more structured view of the potential ethical challenges faced by Web archiving initiatives. It tries to draft or inspire an initial research roadmap, identifying areas of questions that should be addressed in addition to the current research being performed by the community to ensure that Web archives may unfold their potential to the fullest extent in a way that does not conflict with society's valid concerns about privacy and ethically correct usage of large volumes of accumulated data.

The remainder of this paper is organized as follows. Section 2 identifies some of the core challenges that may give rise to ethical concerns, both regarding the collection as well as, specifically, the usage of the information collected. Section 3 identifies a first set of research questions and points out directions that require coordinated investigation in order to pro-actively embrace the concerns identified. Section 4 reports on experiments on the potential of identifying public content and private user comments in Web pages, both on a page level as well as on a paragraph level. Section 5, finally, sums up the main lessons learned, and calls for pro-active discussion and research on these issues from all actors involved in Web archiving endeavors.

2. ETHICAL ISSUES IN WEB ARCHIVING

Web archiving is an important endeavor to safeguard large and crucial parts of our heritage. Yet, there are ethical questions arising from the activities of collecting material from the Web. While legal issues such as copyright can be "simply" solved by corresponding legislation, the situation is different with ethical questions that should merit serious considerations by anybody involved in such activities. In this section we thus want to provide a brief overview of some of the most prominent ethical issues touched upon by Web archiving initiatives. This is neither a complete coverage of all issues, nor an in-depth ethical treatise of the matters involved, but merely meant to raise some questions and concerns that we should be willing to address. The goal is not to stop Web archiving initiatives, but to raise awareness for these issues, providing ethical and technical frameworks to ensure success and widest possible acceptance of Web archiving initiatives. The ultimate goal, in this respect, is to allow Web archives to be open for access and usage by the public at large while minimizing the potential for abuse.

A bit over-simplifying, there are 3 core assumptions made by most Web archiving initiatives justifying their moral right and need to collect the information available on the Web:

1) The Web constitutes a new form of publishing. It is made up of publications that are otherwise not taken care of, and that would be lost if no collection and archiving solution like in place for conventional publications were devised.
2) The Web's ephemeral nature is a deficiency of the medium, incurred by the distributed, unmanaged nature of the system, that could and should be overcome by a more centrally managed archive.
3) The Web archive is merely a collection of material that is freely available anyway.

All three assumptions, while they definitely are true for many parts of the Web, may be contested on a general scale.

2.1 The Web as a publication medium

Let us take a look at the definition of the concept of Internet publications and authorship. It is true that the Web constitutes a new publication venue for many old and new authors, generating new genres such as e-novels, collaborative authoring, on-line journals, and on-line art projects.

Yet, it is highly questionable, whether everybody who creates a homepage; who adds a comment in a discussion forum; who puts up some pictures of last week's outing with friends on a photo sharing Website; or who registers his or her name for a seminar, with the list of participants subsequently being available on the Internet mostly for participants to download; or all pupils who have the assignment of contributing to their school's Web page; whether all of these persons may be considered -- and do see themselves -- as authors in the traditional sense, and consider the process of distributing information to sometimes very limited target groups as publishing? Is anything that children or teenagers shout out to the Web to be preserved and accessible in eternity –

or at least by their future employers? Can we deem all persons who are creating content on the Web sufficiently computer-literate as well as forward-looking to understand the consequences? Do we, or rather: does society, actually want the Web to function that way?

Should we not rather consider Web pages and postings sometimes more like a private communication in public, like a group of friends discussing in a metro train -- conversations that are not meant to be broadcast via the public TV network as well as archived and indexed for everybody to search through? Isn't the Web as much a communication medium as it is a publication medium?

While the act of publication required some considerable effort in the traditional world, and in that procedure ensured that the persons undergoing it were aware that their work was about to be published, this no longer is true in the Web environment. Are all of those "New Authors" publishing blog diaries, their vacation pictures, and so on really aware of the fact that they are publishing them?

2.2 The "Ephemerality-fault" of the Web

Let us now briefly address the second motivation for Web archives, namely the mission to correct a "design error" of the World Wide Web, i.e. its ephemeral nature. Web publications, according to different estimates, have an average life span of a couple of days to a few weeks.

If we do not provide a means to counter this deficiency, everything that people created on the Web, that constitutes important social, cultural, economic or scientific value, will be lost, as the New Authors on the eb do not have the means to provide persistency for their intellectual input in this technologically hostile environment.

Yet, is it really, that the "New Authors" -- if we accept them for a moment to be such -- produce their contributions with the prospect of having it archived? Basically, this again brings us back to the concept of what constitutes a publication. Yet, now we do this with the added dimension that even if we accept the Web as a new publication medium, we have to question whether it is not its ephemeral nature that is characteristic of it, and that is specifically used by authors to create content specifically designed for such a setting. Many artists creating installations and performances do their best to make their creations one-time events that are not to be conserved. Similarly, many documents may be placed on the Web because it is ephemeral, because they are meant as a temporary statement; a comment, that is never intended to be captured and maintained for eternity. Is there a justifiable reason for providing such forms of non-persistent comments, similar to oral statements, in written form?

If we accept that the Web has certain ephemeral characteristics, is it conceivable that some "authors" decide to use the Web as a medium because of this characteristic? This definitely may apply to the world of arts – but probably also to much more mundane and conventional settings. When applying for a job applicants usually decide to optimize their CV towards their current interests and the job they are currently applying for. Rather than sending the application directly, they may find themselves forced by social standards to put this information on their private homepage – which does not constitute a major issue in itself, as they can decide to adapt it whenever we deem fit. It may also be used as a temporal outlet to communicate certain spontaneous comments – meant to be available only for exactly the limited period of time that one created them for. Archiving and aggregating this information across time significantly changes this ephemeral capacity of the medium

2.3 The Archive as copy of free Web material

A third issue that has ethical implications, yet is easily overlooked by Web archiving initiatives is the fact that a Web archive is more than just a collection of individual Web pages. Is it not that the compilation of many Web pages on a certain subject across longer periods of time reveals a completely new point of view on the subject? While this principle also applies to conventional archives, it is the ease (and the fact that this probably will constitute a dominant means of utilization of Web archives) by which these kind of compilations can be created and searched that makes Web archives different: any query, be it for a person's name, a specific topic, or a combination thereof, will produce as a result a compendium of search results, sorted by relevance, date, or any other means, and providing a - most probably not beautifully edited, un-reflected, but nevertheless -- new representation of the individual pieced of information. This also leads to the question, in how far Web archives are different from conventional search engines, and whether unique policies need to be devised to govern them.

According to a recent study by the Federal Union of German consultants (Bunderverband Deutscher Unternehmebsberater, BDU[1]) 28% of human resource managers are already researching an applicant's presence on the Internet. This is currently limited to their present time, not providing a life time history back to school-day views on all kind of issues -- yet. This has been possible even before the days of the Internet, but it required a much higher effort that limited that kind of research to very specific positions, and having trained specialists perform it.

While the reply, that such a complete picture may be closer to the truth than a specifically designed representation at a given point in time may be -- if well-researched -- true, we may still ponder over whether we want that kind of truth (not to mention the potential effects of ill-researched background checks). This has resulted in the creation of specific person search engines in the Internet that scan Blogs, social networks, podcasts, and the Web in general, to identify and compile information on people. Blinx[2], for example, uses multimedia mining to analyze videos and podcasts to identify people's names using speech recognition software. Other specialized search engines are Maltego[3], PeekYou[4], Pipl[5], Spock[6], Stalkerati[7], Wink[8], Yasni[9], YoName[10],

[1] http://www.bdu.de

[2] http://www.blinkx.com

[3] http://www.paterva.com/web/Maltego

[4] http://www.peekyou.com

and ZoomInfo[11], to provide just a few examples of search engines using specialized technology available to us today – with significantly more advanced services to be expected in the near future.

The inclusion of Internet research in the creation of human resource profiling has given rise to initiatives countering these activities by offering reputation management services. These aim at creating optimized personal profiles and linking structures in popular Web 2.0 platforms such as Flicker, Facebook, and others, as contenders to this trend.

There are myriads of other issues that merit further investigation. Given the short period of time that this medium exists, many aspects of its use have not evolved yet into clearly specified and commonly accepted patterns of usage. This should by no means block Web archiving initiatives from pursuing an important task. We might, however, decide to include investigations on certain ethical aspects into the corresponding projects, trying to provide technical, managerial and legal solutions not only for the creation and maintenance of Web archives, but also for proper usage.

Solutions may encompass automatic means to analyze and identify the types/genres of Web pages made accessible (can we separate discussion forums from publications?), types and number of searches permitted for specific concepts, and others. These may range from simple heuristics to highly sophisticated techniques, blocking names, dates, Web pages of a certain genre from being either harvested or displayed.

They may require certain procedural components, such as opt-in and opt-out regulations, temporal access limitations and other legal restrictions. Privacy is an important aspect to many people's lives, that we should address with due diligence.

Most Web archiving initiatives are aware of this and thus do not provide full access to their collections at this point in time, or do have provisions in place to allow users to have their material excluded or removed from the Web archive. However, more sophisticated approaches may provide more flexible and more comprehensive solutions while requiring less knowledge on the existence and ways of operation of specific Web archives by the users. While this definitely will not make Web archiving any easier, it may well be necessary in order to make the resulting archives successful and appreciated.

[5] http://www.pipl.com

[6] http://www.spock.com

[7] http://www.stalkerati.de

[8] http://www.wink.com

[9] http://www.yasni.de

[10] http://www.yoname.com

[11] http://www.zoominfo.com

3. TOWARDS ETHICALLY ACCEPTABLE WEB ARCHIVES

Web archives contain and represent an important aspect of our modern cultural heritage. However, due to the aggregation of information across time, as well as the fact that publishing on the Web is perceived in a different way by many "authors" than publishing in conventional media, Web archives may also contain a lot of information that probably is not intended to be accumulated by the people who created it, shared and searched in the way a Web archive will allow people to do now or in the future, using more advanced search technology. In a nutshell, access to Web archives does raise questions that merit closer analysis. These ethical concerns rightfully force many Web archives to be locked away from public access, rendering the potential value of the information collected next to nonexistent.

In order to unlock the value of Web archives, means need to be found that allow providing access to the content in a Web archive that does not conflict with ethical or jurisdictional concerns. This will require significant research efforts along several directions, some of the most immediate probably being to questions such as

(1) what are the ethical constraints, and how they can be more precisely defined or formalized,
(2) in how far technological solutions such as query analysis, machine learning and data mining can help in identifying potentially harmful queries, potentially incriminating content on Web pages, or combinations thereof,
(3) which approaches potentially malicious users of Web archives might employ to obtain information that should not be provided by privacy-protecting archives, and
(4) how legal regulations might be formulated in order to allow (partial) access to Web archive content in a save, ethically correct, and useful manner

3.1 Ethical implications of Web Archives

In cooperation between Web archivists, data mining experts, as well as specialists on privacy issues and ethics the potential risks stemming from access to the accumulated information stored in Web archives need to be explored. Risk scenarios, and the implications of certain types of information that can be obtained from Web archives have to be investigated in order to be able to as precisely as possible formalize potential threats. Only a precise understanding of threats or concerns that access to Web archive content in various forms may constitute will allow us to develop suitable strategies to mitigate these risks.

3.2 Analysis of Web Archive Content

Different types of content in Web archives have different degrees of sensitivity with respect to privacy issues.

We thus need to identify, in how far techniques like data mining and machine learning in combination with natural language processing techniques, etc. can be employed to identify potentially sensitive bits of information by capturing the underlying semantics of the information at hand. This can be addressed at several levels of complexity, ranging from a classification of sites or pages into different categories of sensitivity (company homepages vs. Blogs or private homepages, sites with illegal or offensive content), via analyzing various parts of information on a single page (news article vs. discussion items following it), up to detailed analysis of certain parts of text (birth dates, offensive words, names, or assumptions being made). This information could be utilized to either block out certain information, provide access to it in a non-comprehensive manner, or limiting access by different other means.

3.3 Query Analysis

In order to identify potentially ethically questionable usages of Web archives, users querying Web archive content may be required to identify themselves, be it via limited access in reading rooms, or by logging into the archive via secured sessions. This will enable archives to track their searching behaviour (which may, in turn, give rise to another level of ethical concerns over the monitoring of users in archives). This information can be utilized to identify potentially unethical utilization of Web archive content. One might differentiate between users trying to collect background information on health conditions of potential applicants for open positions, as opposed to somebody researching his or her family tree. Similar to the complete tracker problem in databases certain sequences of queries might be identifiable that may lead to certain parts of the Web archive being blocked from access or filtered.

3.4 The Malicious Web Archive User

In order to better understand the techniques that might be used by potentially malicious users of Web archives, and the types of information they might be able to gleam from the archive, one method would be to try out various techniques that might be utilized by malicious users. By trying to identify means of abusing the wealth of information collected, archives may be able to better understand the threats and the techniques employed to do so, which, in turn, will provide valuable clues to protect against them.

3.5 Access Policies

In order to be able to grant access to eb archives, legal frameworks and more fine-grained access policies must be put in place that regulate the responsibilities of both providers as well as users of Web archive information. These policies need to be based both on a common understanding of precisely defined ethical risks as well as the technological possibilities if Web archives are to be used. Defining access policies as well as means to ensure that they are being followed will allow access to Web archive information in such a way that the privacy of individuals is preserved. Still, the risks posed by such a limitation of access to information, i.e. “censorship” may not be neglected and require equally thorough consideration [4].

Access regulations are further complicated by the transnational linkage existing on the Web, while many existing Web archiving initiatives are limited o specific national or topical sectors. This will require international policies to be formulated in order to to allow Web archives to serve their purpose – while at the same time protecting the legitimate interests of citizens concerning their data.

4. IDENTIFYING PRIVACY-SENSITIVE CONTENT IN WEB ARCHIVES

This section reports on some initial experiments trying to identify sections in Web documents that are either official/public information or private user comments. The idea behind this strategy is to be able to subsequently perform more in-depth analyses of these segments in order to decide whether appropriate measures, such as excluding the content from the full-text searchable index, or blacking out specific text segments (names, phrases, dates) may be advisable when displaying content.

In order to identify elements of Web archive content that may contain potentially privacy-sensitive information, we have launched a series of experiments that analyze the structural as well as content of text elements in Web pages. By extracting numerical descriptors as they are being used in information retrieval, particularly in text classification and genre analysis, we aim at obtaining descriptors that allow a segmentation and classification of text paragraphs into the two categories of private and obviously public information. The primary goal here was less on the actual concept of what types of content may be privacy-sensitive, but to identify if – given any such splitting – machine learning models may be employed to automatically learn the underlying concepts and to differentiate them. The figures reported below should thus not be taken as absolute indicators of achievable performance, but rather demonstrate the feasibility of this approach.

4.1 Text corpora

Two different text corpora were used for the experiments. One is the 7-web-genre-corpus [7], a public benchmark corpus that is frequently being used in genre analysis experiments. This corpus consists of 1400 documents (200 per genre), which are grouped into 7 main genres, namely *Blog*, *Personal Homepage*, *E-Shop*, *FAQ*, *Listings*, *Newspaper Frontpage* and *Search Page*. Of these, the first two were defined to contain potentially privacy-sensitive information, whereas the other 5 genres were defined to contain only information that may rightfully be distributed publicly also in an archival setting. For this experiment, all documents were considered as a whole; i.e. entire documents (Web pages) were classified to be public or private.

In a second experimental setting documents from product review sites were collected and analyzed on a paragraph level. Specifically, we collected 50 pages from the “Top 100 Books of

2007" listing of amazon.com. These contain both descriptions of the actual books, as well as comments by private users. The former parts of the pages were labelled as public information, while the latter were considered private. Furthermore, 20 pages describing video games were selected, again from amazon.com, and treated in the same way. The reason for analyzing these two categories lies in the significantly different styles both of the descriptions as well as the comments made by users for these two types of products.

4.2 Features

A range of different features were extracted from the text corpus, using a normal full-term indexing system, as well as a part-of.-speech tagger (Gate, [8]) or readability indices ([1,10]). The list of features used is given in Table 1.

Table 1. Features used for indexing documents

Category	Features
Text statistics (averages)	# words, +syllables, word-length
Tokens (rel. frequ.)	word-, symbol-, space-, number-, punctuation- control-tokens
Readability Indices	Flesh Reading Ease, Flesh-Kincaid Grade Level
Look-ups (rel. frequ.)	places, time/date, person names, currencies, addresses, company names, abbreviations, telephone/fax numbers
Part-of-Speech (rel. frequ.)	nouns, verbs, adjectives, adverbs, modal verbs, conjugations, pronouns, personal pronouns, articles, prepositions, exclamations, list elements
Presentation (rel. frequ.)	links, headings, paragraphs, text formatting, lists, tables, graphics, frames, forms, multimedia elements

For the second test corpus, several different approaches for segmenting the Web pages into paragraphs were used. These were based both on formatting indicators in HTML, as well as on significant changes in the word types being used in subsequent lines of text.

4.3 Classification results

In order to automatically classify documents and paragraphs into either the public or the private category, we used three classifiers, namely Naive Bayes, SVM, as well as a knn-classifier, as implemented in the WEKA machine learning package [13]. Validation of results was based on 10-fold cross-validation.

The classification results for the 7-web-genre corpus is given in Table 2 for the individual classes, as well as overall for the *private* category using SVM.

Table 2. Classification results using SVM

Genre	Precision	Recall	F2
Blog	93,17	87,22	90,01
Personal Homepage	89,59	48,33	62,79
Private	91,38	67,78	76,40

However, the results above constitute basically just a simplified challenge in terms of genre classification [7, 9]. A more realistic setting is provided in the second corpus, where individual paragraphs are being analyzed and categorized accordingly. As the second corpus is rather specific, in the sense that it only contains documents from one specific source, no specialized model was built from this corpus. Building a specialized model would obviously have resulted in significantly higher performance, but would limit the usefulness of the results in the sense that they would be biased to this specific corpus, which, in turn, would require similarly specific models for any other type of pages. Rather, we used the model obtained from the generic 7-web-genre corpus above directly to classify the individual paragraphs. As segmentation boundaries between paragraphs were not always identified correctly by the automatic segmentation methods implemented, we evaluate results on a character level. That is, we analyze the percentage of characters that were assigned to the right category of private or public information. Results are provided in Table 3 for the Naive Bayes classifier. In order to highlight the sensitivity of the classifier towards the specific writing styles, we identified two subsets of readers' comments for the top-100-books corpus: The pages contain both professional reviews, which we may consider public comments similar to objective, public texts, as well as "normal" users' comments, which we consider as potentially private data in a separate evaluation. Results for this sub-analysis are listed the top-100-books-b. It shows that, in addition to splitting comments from the book descriptions, the classifier also is able to really analyze the style of the document written, distinguishing "official" comments from more personal ones.

Table 3. Classification results, paragraph level (Naive Bayes)

Corpus	Precision	Recall	F2
top-100-books	77,91	84,07	80,87
top-100-books-b	78,27	85,75	81,84
video games	87,66	96,02	91,65

4.4 Discussion

The results above show that the proposed approach is able to automatically identify professional/objective statements from private user comments with significant accuracy. While this alone cannot and should not be the basis for hiding or disclosing information in a Web archive, it may serve as a baseline for further analysis, focusing more in-depth analytical efforts on these segments. These may reveal the type of comment made, or the type of document it represents (as e.g. perfectly public Web pages may be written in an informal style to mimic a colloquial structure). The nature of the information revealed may assist in deciding whether specific protective measures should be taken.

These may include double-checking whether the identity of the person making that comment is revealed, whether sensitive information is revealed that should be blocked when displaying the page, and others.

5. CONCLUSIONS

The creation of Web archives constitutes an important task in ensuring that a valuable aspect of our cultural, social and scientific heritage is being preserved. At the same time, given the wealth of information that these archive will contain, and the powerful techniques that we will have available for searching and analyzing them, we are faced with a number of challenges in order to ensure that the archives we are creating do not pose a threat to the privacy of people, or give rise to other ethical concerns.

Most Web archiving initiatives are aware today, that a number of ethical issues exist in the context of creating and providing access to their holdings. A range of approaches have thus been deployed to ensure that they contain only information if the creators do not disagree with their content being archived, or that they are being used in the right way. Some archives obey robots exclusion protocols, or remove pages on users' requests. Others block access to their holdings, or provide it only on a very limited basis.

The methods and experiments described in this paper are not intended to serve as a direct filter in any access module to Web archive, but to investigate the feasibility of identifying specific types of information. We are currently expanding the work reported above in two main directions. One goal is to further generalize the principles identified above and to investigate in how far the performance figures obtained generalize across different types of Web pages. In addition to that we are studying a hierarchical approach to provide a background bias for individual text segments by analyzing the semantics and style at different levels, ranging from text paragraphs via pages, sets of pages, up to an entire domain.

In a second line of research we are attempting to create a more fine-granular semantic analysis of text segments. This should provide us with a better understanding of the nature of information being presented, allowing to specifically identify offensive statements, or the disclosure of sensitive information.

In order to ensure that we may create and exploit Web archives to their fullest extent and to the benefits of society at large - now and in the future - coordinated and concentrated efforts are required to identify and obtain a better understanding of the real challenges that we are faced with, both in the creation as well as in the usage of these collections. We further need to identify legal, organizational, and technical means to provide a solid basis for the operation of our archives. At the same time, we also need to analyze the potential dangers that these means to ensure the privacy of data constitute. They may be intended to make Web archives meet the ethical expectations of society, but may prove counterproductive both due to intended as well as unexpected effects of rules and regulations or technical imperfections of the solutions and heuristics applied. While the challenges are significant, and both technical as well as organizational solutions available today solve those only to a limited degree, it seems feasible to develop suitable solutions, if we embark on a pro-active role in identifying, naming, and addressing those issues.

6. REFERENCES

[1] McCallum, Douglas R. und James L. Peterson: Computer-based readability indexes. In: ACM 82: Proceedings of the ACM '82 conference, pp. 44–48, New York, NY, USA, 1982. ACM.

[2] Foot, Kirsten A., Schneider, Steven M. and Dougherty, Meghan: Ethics of/in Web Archiving. Presentation given at the 2003 Conference of the Association of Internet Researchers, October 10-19 2003, Toronto, Canada.

[3] Franke, Mark S., Siang, Sanyin: Ethical and Leagl Aspects of Human Subjects Research on the Internet. Report of the AAAS and OPRR Workshop on Ethical and Legal Aspects of Human Subjects Research in Cyberspace. Tech Report, American Association for the Advancement of Science, Washington, DC. November 1999.

[4] Grimmelmann, James: Don't Censor Search. In 117 Yale Law Journal – Pocket Part 48, 2007 http://thepocketpart.org/2007/09/08/grimmelmann.html

[5] Kavcic-Colic, Alenka: Archiving the Web – Some legal aspects. In: Proceedings of the 68th IFLA Council and General Conference, August 18-24 2002

[6] Pasquale, Frank: Asterisk Revisited: Debating a Right of Reply on Search Results. In: Journal of Business and Technology Law, 2008, to appear.

[7] Santini, Marina: Automatic Identification of Genre in Web Pages. PhD thesis, Univ. of Brighton, Brighton, UK. 2007.

[8] Sheffield, University of: GATE, A General Architecture for Text Engineering. Website, http://gate.ac.uk.

[9] Stamatatos, E., N. Fakotakis und G. Kokkinakis: Text genre detection using common word frequencies. In 18th International Conf. on Computational Linguistics, 2000.

[10] Talburt, John: The Flesch index: An easily programmable readability analysis algorithm. In: SIGDOC '85: Proc. of the 4th Annual International Conf. on Systems Documentation, pp. 114–122, New York, NY, USA, 1985. ACM.

[11] Tofel, Brad: "Wayback" for Accessing Web Archives. In: Proceedings of the 7th International Web Archiving Workshop, Juni 2007, Vancouver, Kanada.

[12] Vaidya Jaideep, Clifton Chris, and Zhu Michael: Privacy Preserving Data Mining, Springer, 2006.

[13] Witten, Ian H., Eibe Frank, Len Trigg, Mark Hall, Geoffrey Holmes and Sally Jo Cunningham: Weka: Practical Machine Learning Tools and Techniques with Java Implementations. In: Proceedings of the ICONIP/ANZIIS/ANNES'99 Workshop on Emerging Knowledge Engineering and Connectionist-Based Information Systems, pp 192–196, 1999. Dunedin, New Zealand.