

Social Considerations in Choosing Digital Archiving Technologies

Chao-Kuei Hung and Yen-Liang Shen

Information Management Department
Chaoyang University of Technology
Wufeng, Taichung, Taiwan

Abstract. Social considerations are important in choosing digital archive technologies. Inappropriate choices that encourage vendor lock-ins may result in software monopoly and illegal copying of software, and unnecessarily widen digital divide. It is recommended that archive maintainers standardize on interface instead of standardizing on software, search for existing, international standards on open interfaces, and collaborate with instead of forking from such existing standards. Misconceptions about the insurance of openness that XML brings need be cleared. As well, distinction between openness of interface and openness of source code need be made. Public awareness of consumer right issues and other social issues in digital technology need be raised by integrating these elements into existing computer education as well as into government propaganda against illegal copying of software.

Index Terms — **digital archives, social, open interface, open file format, proprietary file format, open source software**

1 Why are Social Considerations Important?

The digital technology is evolving so fast that the field of archiving is faced with bewildering and very often mutually conflicting choices of emerging technical solutions. Worse yet, some choices may be dressed up and advertised with emphasis on ease of use, attractive appearance both in the final digital documents and in the tools themselves, etc. while only revealing its conflict with the long term objectives of digital archiving at a much later time when the adopting organization find it too late and too costly to switch to other alternatives. In this paper we look at the problem of choosing technologies for digital archives *not* from the viewpoint of technical merits but rather from social perspectives.

The emphasis on the social perspective is justified by several natures of digital archives involving the general public.

Many digital archives created and/or maintained by the government are meant for public access and dissemination. In fact with the few exceptions concerning national security and citizen privacy, it is probably quite reasonable for the tax-payers to request access to most of the digital archives whose construction receive major funding from the government. There is a trend towards releasing the contents of digital archives in some form of creative commons license [1],

not only voluntarily by contributors of such projects as Project Gutenberg [2], the GNU free software project [3], wikipedia [4], the MIT open courseware [5], etc., but also being actively advocated and requested by such communities as the public library of science [6] and the science commons [7]. It is irrelevant whether certain archive is currently released under such licenses, nor is it relevant whether archive holders and/or the content authors agree or disagree with such a trend. The point is to be prepared for such a strong possibility in view of its relative success. Such success is especially eminent in the free software front for example, considering how little commercial interest and sponsorship there was at the movement's infancy.

Regardless of the licenses of the *contents* in a digital archive, it is almost always beneficial to the content owners, the archive maintainers, and the prospective users of the contents alike, that the *existence of the contents* and the *abstract* be easily searchable and accessible. In other words, we would like the *meta data* and *summary information* of a digital archive to reach the public with as few obstacles as possible, either with commercial or public service motivations.

When the meta data and/or a piece of the contents reaches a citizen, s/he is required to use some software to display and/or further manipulate it. If specific software product from a certain vendor is required for such purposes, encouragement of public dissemination of the digital archive necessarily translates into encouragement of massive adoption of such software by the public. This can be a major cause of illegal software copying if the software has a high price. Furthermore, it necessarily results in software monopoly regardless of the price of the software.

Yet another social factor at play is at the archive content creation end. If we would like the content to reach the audience in a format without the aforementioned problems, we had better ensure that the authors provide the contents in such a format in the first place, or else the mediating/converting cost will be prohibitively high for the archive maintainers. Yet there may not be easy ways for the archive holders to stipulate and enforce the format requirements due to the potential diverse backgrounds of the authors, and due to the inappropriateness of the large amount of knowledge being conveyed as merely administrative requirements. Take the National Digital Archive Project [8] for an example. There is great difficulty, especially from administrative perspective, to enforce upon the content authors, who are privileged scholars, such basic requirement as web page compliance to w3c recommendations. If we can indeed reach a consensus on a set of better archiving practices and choices of technologies, we had better communicate the relevant parts of such set to the general public, of whom some may be potential authors. Thus there is great need for some of the consensus to be integrated into the education system.

We will discuss the above issues in the following sections, clearing certain popular misconceptions on the way.

2 Proprietary File Formats: A Technical Issue Causing Social Problems

In view of the public nature and long life of digital archives, especially those maintained by the government, we ask the following important questions:

- Is a citizen forced to install a specific program provided by a specific software vendor on his/her computer?
- Will the option be open for the archive maintainer to mass-migrate the digital archive into some other format in the future?
- Will the reader/writer software be available a century from now?

Several social problems are created when digital archives are published in a format that requires the installation of a specific program provided by a specific software vendor. Firstly, the archive maintainer in effect helps promote the monopoly of the software vendor. The situation becomes even more severe when the vendor's software interoperates with only certain other software, a certain operating system for example, and/or it runs on only certain hardware. Eventually archive-maintainer-encouraged monopoly of one piece of software will induce monopoly in other sectors of the software industry and even monopoly in other industries. [9] Secondly, citizen privacy may be intruded. Once the software vendor successfully seeds its software in every citizen's computer, it opens up a wide varieties of possibilities for it to collect citizens' identity information, browsing habits, etc. The intention of the software vendor might be consumer preferences and the like, but the result might be the leak of citizens' personal information, involuntarily and unknowingly, or even death. [10] The legal and social consequences could be rather severe, to the content users as well as to the content authors and archive maintainers. [11]

We emphasize that all these statements are made without reference to the price of the software. Even if the price is zero, the above social problems still arise. If the software is not free of charge, however, such choice of digital archiving technology further encourages illegal copying, and unnecessarily enlarges the digital divide considering the difficulty of acquiring such software in underdeveloped areas of a country, legally or illegally.

The mass-migration requirement is a possibility that archive maintainers should not deny themselves from the onset. Cryptographic technology evolves, for example. The most popular public-key crypto system of present day may face the attack of the emerging quantum technology, which itself in turn may be used to construct new crypto systems. [12] It is unrealistic to assume that today's technologically best archiving format remains good enough for 20 years. We may even want to migrate the archive to some other format within a much shorter time frame for political or economical reasons. The GIF patent case, for example, would force many maintainers of digital image archives to pay a hilarious amount of money had there not existed a way to mass-migrate the .gif files to .png files. [13] This example also illustrates the danger of employing a file format which is popular, has gratis reader/writer software, but patented. Here

again the temporary zero price of software is not to be taken as an insurance for comfort as the patent holder may choose to exercise his right at a later, more convenient time.

The above social problems have already manifested as a result of the popular use of proprietary file formats in digital archives. There may be even more disastrous effects yet to come from a longer point of view. After all, digital technology is still of young age. In glaring contrast to the life span of shorter than a decade for most of today’s software, digital archives are meant to persist centuries or even longer. Would future generations be able to read our archives encrypted in a proprietary file format? The founders of Project Gutenberg had the foresight of deciding to use the plain text format as the basis for storage, thus making its content survive more than 30 years. [2] Had they mandated the use of then-popular software such as word-perfect or page maker, the content would have been much less accessible today to say the least.

3 Standardize on Interface, not on Software

A very deeply held mis-conception is to “standardize” on a certain software, i.e., to stipulate use of a certain set of software in an archiving system. [14] For example, it is quite common for the conference organizers to require submission of papers be prepared with a certain version of Microsoft Word. A presently less popular but rapidly rising trend is to require the use of Open Office. [15] Such mistaken concept of “sameness” for “compatibility” would look obviously absurd if we make an analogy using the telecommunication technology. Is one required to use the same make or even the same model of mobile phone when answering a call from Nokia 8210? Is it reasonable for an old style telephone to refuse answering at all simply because the call comes from a new style telephone with caller-id capability? *Requiring sameness* – the use of a specific software program – to read/write a document or to interact with a certain part of the digital archive, is precisely the indication of *lack of compatibility*. “Standardizing” on software is the cause of the many social problems discussed earlier.

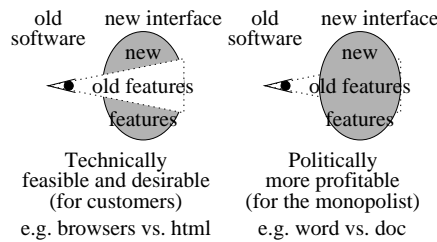


Fig. 1. Old software should be able to see old features such as simple texts and simple graphics in a document stored in a new file format (interface)

We echo Fair’s view that the choice of archiving technologies should standardize on *open interfaces* instead of standardizing on *popular software*. Take word processing for example, the Open Office is a preferable alternative to the Microsoft Office *not because of software quality nor because of software price* but because the specification of the *sxw* file format is publicly available for any interested party, including Microsoft, to implement readers and writers.

4 In Search of Existing Open Interfaces

File format is only one example of interface in software systems. Client-server architectures depend on *communication protocols*; application programs request services from operating systems through *application programming interfaces*; operating systems drive hardware through *Hardware Programming Interfaces*. [16] *It is essential that the openness of all interfaces, including but not limited to file formats, be a fundamental requirement in all choices of digital archiving technologies in order to avoid vendor lock-ins and the ensuing social problems as discussed earlier.* When choosing an archiving technology to deploy in a large scale, one must not forget that the following questions are far more important than the convenience and flashy features of the software. Can the new system be decomposed into components? Do the components talk to each other through open interfaces?

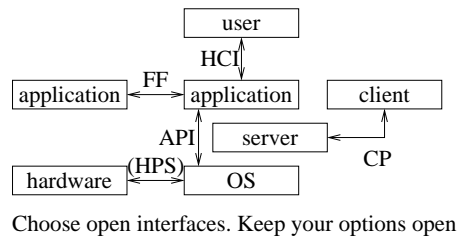


Fig. 2. Software *can be and should be* decomposed into components, especially in defense of vendor lock-ins

With all the importance, obviousness of necessity, and far-reaching effects of open interface, it simply did not receive due attention in the past. Apparently the small but vocal FLOSS (Free/Libre/Open Source Software) community are among the first few to notice the proprietary interface problem. Dennis in his letter to the editor of the Linux Weekly News has a very elucidating proposal on how to remedy the Microsoft monopoly phenomenon with fairness both to the society and to Microsoft. [21] Moreover, the emphasis is not so much to punish Microsoft as it is to prevent future monopoly. This should draw the attention of archive maintainers worrying on the possibility of deciding on a wrong technology that will only manifest as yet another monopoly a few decades from now.

In essence he proposes to open up the specification with an open source, cross-platform, ANSI-compliant reference implementation for the interface in question, be it a file format, an application programming interface, or a communication protocol. The precise definition of open interfaces, however, does not seem to exist until recently, and mostly centered around open file formats alone. [17] To the best of our knowledge, the only serious attempt to define the openness of file format comes from the Texas Open Source Initiative Mailing List. [22] And it probably is not until Hancock, one of the authors, published an article appealing to the FLOSS community, that it became better known. [23] We believe, however, that it deserves the attention of a much larger audience than the FLOSS community.

It is nevertheless possible to summarize a few key features in deciding whether an interface is open or not.

- non-discrimination against any party
- public availability of complete technical specifications
- public availability of FLOSS reference implementation
- free of patent and free of any other legal threats

In addition, it is often desirable, though not always necessary, that an interface such as a file format be *transparent*. [24]

Adobe's pdf, for example, satisfies the first two but not the last two requirements. There exist gratis readers from Adobe for many different OS platforms but they are not open source and they are not scriptable. The FLOSS alternative xpdf and related programs are scriptable but does not display correctly pdf files of newer versions (1.4 or above) containing Chinese characters. Mass-migration from pdf to some other formats will likely become a potential trouble when the need arises. Moreover, pdf is *opaque* as opposed to being transparent, meaning that re-use is inconvenient or sometimes impossible. It is thus an unsettled debate whether the pdf format is a good choice for digital archives, or for word processing in general. The general attitude towards pdf, among people who are aware of the proprietary file format problem, is conditional acceptance with reservation. [25] [26] [27]

It is also important for the archive maintainers looking for open interface solutions to cooperate with existing, international efforts instead of creating one's own variants. Unjustified forking, or divergence from existing standards, are likely to become isolated efforts which will not benefit from further advances and improvements of the existing, international society. For example, it would be wise to follow w3c recommendations in standardizing web pages [28], and to follow the Open Archive Initiatives [29] in standardizing metadata formats and interchanges between archive subsystems or between archives. Should there be any need to diverge from the existing standards, it would be to the benefit of the archive maintainer along with its authors and users as well as of the rest of the world, for the archive maintainer to submit requests to adopt certain changes in an international framework. Moreover, it is advisable to make transparent the process of the development of an open interface, and actively encourage public participation following the RFC (Request For Comment) practice. [30]

Specifically, the following places are nice starting points to look for existing, international open standards.

- The Internet Engineering Task Force: <http://www.ietf.org/>
- World Wide Web Consortium: <http://www.w3.org/>
- Organization for the Advancement of Structured Information Standards (OASIS): <http://www.oasis-open.org/>

However, it must be emphasized that even in such mostly openness-friendly web sites one may still occasionally find some “standards” under serious debate as to whether they are open enough. OASIS’s new and controversial policy allowing patents in “standards”, for example, has recently received much criticism and even boycott. [31]

In view of the lack of a precise definition of open interfaces, we propose a very simplistic test that approximates openness. Try to mix software components from competitive vendors *and from FLOSS*, and watch for potential vendor lock-in attempts when one piece of software does not interact correctly with another piece from a competing vendor or from the FLOSS alternative. [16] In such cases, require both sides to support open formats, not the less popular one to support the more popular one. Software vendors advocating use of their proprietary formats most often vigorously support *importing* from other formats, open or proprietary, while disregard or support half-heartedly *exporting* into other formats. [17] They then confuse the consumers as if their product had a better support than other software. Reverse engineering to achieve exporting may be technically feasible but always a waste of public resources. In addition, the vendor in question may even actively prohibit exporting of their formats with legal threats such as patent [18], the 1201 anti-circumvention provisions of DMCA [19], and/or UCITA [20].

In summary, there is really a very strong need for the industrial, open source, legal, government, and academic communities to come up together with an exact definition of open interface. It will have to encompass not only such interfaces as mentioned above, but also emerging interfaces arising from new technologies and practices such as client-server architecture, application service provider, etc.

We finally note that the current trend of converging on XML is an encouraging one, but *use of XML alone does not ensure openness of the file format or protocol*. It is also essential that the document type definition (DTD) or schema be open. Archive maintainers need be aware of a subtle trend of certain vendors touting openness of their XML format when in fact the DTD/schema or the embedded binary objects are proprietary.

5 Three Levels of Freedom: Open Source Software, Open Interface, and Opiumware

Archive maintainers also need be aware of the distinction between open interface and open source software.

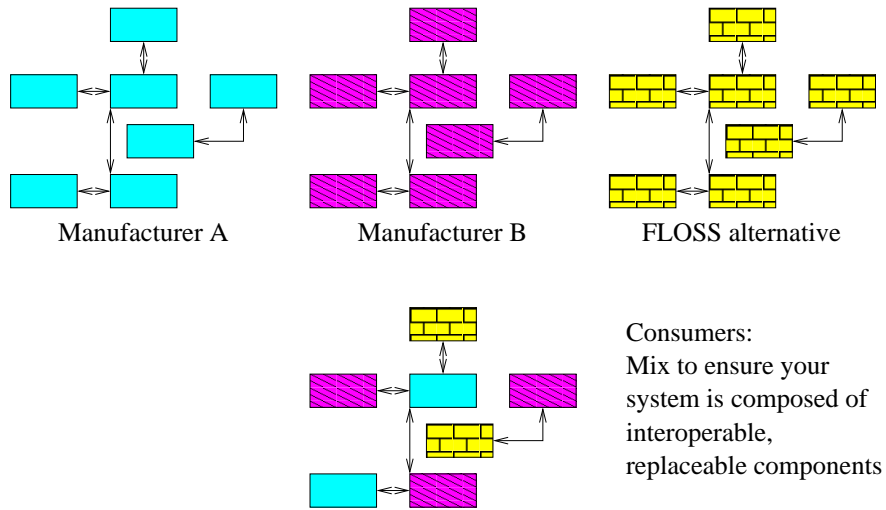


Fig. 3. Mix components from competitive vendors and from FLOSS alternatives to ensure interoperability

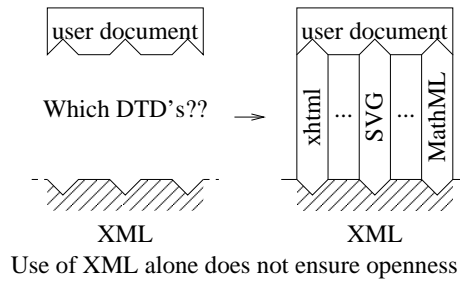


Fig. 4. Use of XML alone does *not* necessarily ensure openness of the file format or protocol

Open source software almost always support open interfaces by default since the inner working of the software is completely visible to anyone. Also, open interfaces almost always come with open source reference implementations. It is another issue whether such open interfaces provided by any specific open source software are the best choices for archive maintainers, when popularity and other issues are taken into consideration. However, it is for sure always a safe choice since everyone is allowed to copy, study, and modify the software itself. The interface it provides can therefore always be adapted to some other open interface if it turns out not to be the best or the most popular choice.

At the other end of the freedom spectrum, a proprietary interface deprives even the most fundamental freedoms of the users. Even if the users created a file of his own from scratch, it could be difficult for him to convert his very own file into some other formats later on if the file was stored in a proprietary format. In view of its addictive nature, we propose to call it *opiumware* for easier public dissemination of the danger of such software.

We emphasize the existence of the often neglected *middle part of the freedom spectrum*. There exist non-open-source software that supports open interfaces. The files they read/write or the protocols they speak are in formats whose engineering specifications are open, even though the software itself may not be open in terms of providing users the freedom to study, distribute, or just copy the code. For example, Star Office and DreamWeaver, when properly configured and directed, can produce OASIS-compliant sxw files and w3c-compliant html files, respectively. Such software allows fair competition from other vendors or from open source software writers. It leaves the users with the freedom to convert their files into other formats, but not the freedom to study and copy the software itself.

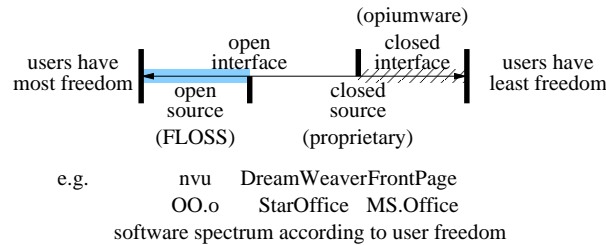


Fig. 5. Three levels in the spectrum of software freedom

An issue emerges in view of the three levels of freedom. Do the archive maintainers enforce use of open source software or simply open interface?

Open source software usually supports open file format better. For example, FLOSS programs nvu, mozilla composer, and OpenOffice.org are among the best W3C-compliant of all web authoring tools, open source or proprietary. Other studies also show that it benefits the authors, maintainers, and users of the archive to make use of open source software for financial, learning investment,

and many other reasons. [32] However, it is difficult to enforce and may sometimes evoke strong objection especially from parties whose commercial interests are adversely affected. Their arguments against requiring use of FLOSS very often divert discussions to issues unrelated to archiving such as the incentives for the vendors to create software. Also, as we mentioned earlier, it is better to standardize on interfaces, not on software.

On the other hand, mandating use of open interfaces is a *much weaker and more lenient requirement*, and therefore has *much stronger and compelling supporting arguments*, such as those found in this paper and many of its references. It has nothing to do with requiring the vendor to give up their copyright on the software. To make a metaphor, we are building a technological Tower of Babel today with the help of many pieces of software. The insistence on open interface is a humble request that everyone be allowed to speak the same language without legal hassles. Any attempt to dissuade its enforcement only discloses further the intention of the advocates to create lock-in and monopoly at the cost of the Tower of Babel and at the cost of the entire society.

In summary, open source is a worthy endeavor, and open interface is a must. Archive maintainers are advised to hold a firm and uncompromized position for open interfaces, to investigate the use of open source software and encourage its adoption but advised against mandatory policies based on software title whether or not it is FLOSS.

6 Conclusion: Computer Literacy Need More Social Elements

As the readers may have noticed, there is a huge chasm between the technical community and the general public in terms of computer literacy, and in-between lie the largely ignored concepts about file formats, open interfaces, open sources, etc. Whether these concepts get public attention and dissemination has a strong influence on social matters such as illegal copying of software, citizen privacy, and monopoly. Yet they are largely ignored by the technical community because their focus is on more technical activities such as reading specifications, writing codes, and inventing new algorithms. And they are hardly present in any of today's computer literacy curriculum for the general public, for whom learning computer is equivalent to selecting items in a user-friendly graphical menu.

Archive maintainers need to face both the technical communities and the general public, and need to be responsible for long term social effects of the chosen digital technologies. They are among those who first see the adverse result of today's inadequate computer curriculum. We propose that archive maintainers join force with the education sector to introduce certain long-ignored important elements into our current computer curriculum for the general public, such as:

- consumer right to interchange software components
- consumer right to space-shift and time-shift files
- inappropriateness of spreading proprietary file format from an ethical and social point of view

- standardizing on interface, not software
- distinction between open interface and open source
- separation of content from presentation
- security by obscurity is no security
- ...

The point about inappropriateness of spreading proprietary file format is especially suitable for inclusion in government propaganda against illegal copying of software. All of the above points need to be conveyed to the public, in plain, non-technical words, and better yet, illustrated with analogies and diagrams. For example, Smith explained proprietary file format using the analogy of a person requesting another person to take notes while the latter refuses to use a commonly known language to do it. [17] We need more intuitive explanations like this one.

These goals may seem far away. And yet it is worth starting now the remedial actions in today's computer education for the general public, even though it is the digital archive maintainers of future generations, not of today, who are more likely to enjoy the full benefits of their fruit. After all, digital archives concern matters of importance *in the long run, and people who are in charge of them are in the best position to see far enough into the future.*

References

1. *Creative Commons [Online]*. Available: <http://creativecommons.org/>
2. M. Hart, et. al. , (1971). *Project Gutenberg [Online]*. Available: <http://www.gutenberg.org/>
3. *The GNU Operating System [Online]*. Available: <http://www.gnu.org/>
4. *Wikipedia, The Free Encyclopedia [Online]*. Available: <http://wikipedia.org/>
5. *MIT OpenCourseWare [Online]*. Available: <http://ocw.mit.edu/>
6. *Public Library of Science [Online]*. Available: <http://www.plos.org/>
7. *Science Commons [Online]*. Available: <http://science.creativecommons.org/>
8. *National Digital Archive Project [Online]*. Available: <http://www.ndap.org.tw/>
9. D. Kegel, *On the remedy phase of the Microsoft antitrust trial [Online]*. Available: <http://www.kegel.com/remedy/>
10. A. Mao, (2003, October). *.doc files reveal your secrets [Online]*. Available: <http://www.lins.fju.edu.tw/mao/works/nodoc.htm> (traditional Chinese)
11. G. Rangwala, (2003, September). *Inquiry into David Kelly's death takes emotional turn as powerful UK communications director Alastair Campbell resigns [Online]*. Available: <http://www.democracynow.org/article.pl?sid=03/09/02/1413225>
12. J. Ford, (1996). *Quantum Cryptography Tutorial [Online]*. Available: <http://www.cs.dartmouth.edu/~jford/crypto.html>
13. P. Sarrazin, (2000, April). *Unisys/CompuServe GIF controversy [Online]*. Available: <http://lpf.ai.mit.edu/Patents/Gif/Gif.html>
14. E. Fair, (1997, October). *Software standards versus protocol standards [Online]*. Available: <http://www.clock.org/~fair/opinion/open-standards.html>
15. *OpenOffice.org [Online]*. Available: <http://www.openoffice.org/>
16. C. Hung, (2001, March). *Consumers: Watch your right to free flow of information [Online]*. Available: <http://www.cyut.edu.tw/~ckhung/a/c010.shtml> (traditional Chinese)

17. I. Smith, (2003, August). *The Rise of Proprietary Formats* [Online]. Available: http://opensource.mimos.my/fosscn2003cd/paper/slides/09_imran_william_smith.pdf
18. (2005, January). *Software Patents in Europe: A Short Overview* [Online]. Available: <http://swpat.ffii.org/log/intro/index.en.html>
19. *Electronic Frontier Foundation: Digital Millennium Copyright Act (DMCA) Archive* [Online]. Available: <http://www.eff.org/IP/DMCA/>
20. *What's Wrong With UCITA?* [Online]. Available: http://www.ucita.com/what_problems.html
21. J. Dennis, (2000, April). *Protocols, APIs and File Format Libraries* [Online]. Available: <http://old.lwn.net/2000/0504/backpage.phtml>
22. S. Baker, T. Hancock, et. al., (2003, March). *Open file format definition* [Online]. Available: <http://www.anansinspaceworks.com/Documentation/BuildImage/Legal/tosi.openformatdef.2003.03.19.html>
23. T. Hancock, (2005, January). *Free file formats and the future of intellectual freedom* [Online]. Available: http://www.freesoftwaremagazine.com/free_issues/issue_01/focus_format_free/
24. *Free Document License* [Online]. Available: <http://www.gnu.org/licenses/fdl.html>
25. M. Chakravarty, (2005, March). *Email Attachments* [Online]. Available: <http://www.cse.unsw.edu.au/~chak/email.html>
26. N. McBurnett, (2004, Feb). *PDF vs HTML* [Online]. Available: <http://bcn.boulder.co.us/~neal/pdf-vs-html.html>
27. A. Ertl, *What is the PDF format good for? Nothing.* [Online]. Available: <http://www.complang.tuwien.ac.at/anton/why-not-pdf.html>
28. D. Raggett. *Clean up your Web pages with HTML TIDY* [Online]. Available: <http://www.w3.org/People/Raggett/tidy/>
29. *Open Archives Initiative* [Online]. Available: <http://www.openarchives.org/>
30. *Request for Comments* [Online]. Available: <http://www.rfc-editor.org/>
31. L. Rosen, B. Perens, et. al., *A Call to Action in OASIS* [Online]. Available: <http://perens.com/Articles/OASIS.html>
32. C. Hung, (2001, August). *Effective Computer Learning Strategies* [Online]. Available: <http://www.cyut.edu.tw/~ckhung/a/c013.shtml> (traditional Chinese)
33. (2004, November). *How Open Can Europe Get?* [Online]. Available: <http://xml.coverpages.org/IDA-0FE-18033.pdf>
34. C. Burstein. *Viewable with Any Browser: Campaign* [Online]. Available: <http://www.anybrowser.org/campaign/>