Journal of Zhejiang University SCIENCE
ISSN 1009-3095
http://www.zju.edu.cn/jzus
E-mail: jzus@zju.edu.cn

Machines as readers: A solution to the copyright problem

SHAMOS Michael I.
(*School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA*)
(*Department of Computer Science, University of Hong Kong, Hong Kong, China*)
E-mail: shamos@cs.cmu.edu
Received Aug. 5, 2005; revision accepted Sept. 10, 2005

Abstract: Copyright and its international complications have presented a significant barrier to the Universal Digital Library (UDL)'s mission to digitize all the published works of mankind and make them available throughout the world. We discuss the effect of existing copyright treaties and various proposals, such as compulsory licensing and the public lending right that would allow access to copyrighted works without requiring permission of their owners. We argue that these schemes are ineffective for purposes of the UDL. Instead, making use of the international consensus that copyright does not protect facts, information or processes, we propose to scan works digitally to extract their intellectual content, and then generate by machine synthetic works that capture this content, and then translate the generated works automatically into multiple languages and distribute them free of copyright restriction.

Key words: Universal Digital Library (UDL), Copyright, Digital rights management, Compulsory license, Berne convention, Public lending right, Synthetic documents, Machine translation
doi:10.1631/jzus.2005.A1179 **Document code:** A **CLC number:** TP391

INTRODUCTION

Everyone associated with digital libraries, and especially participants in the joint China-India-USA Million Book Project, assumes that it will be beneficial to mankind to digitize all the works ever published and make them available over the Internet, whether for fee or otherwise. However, it is difficult to find a clear printed agenda explaining how society will be enriched aside from the obvious advantages of preservation, indexing and efficient digital distribution. I propose that the ultimate impact of a Universal Digital Library (UDL) cannot be realized unless computers are able to read, process, paraphrase and translate its contents. The reasons involve copyright law, human behavior and the limitations of our cognitive power to assimilate information.

It is estimated that 100 million books have been published since Man began writing them. Something over half of them can be found in the combined libraries of the world. As of September 2005, OCLC's WorldCat listed over 57 million records on items spanning the last 3 000 years. While 100 million books is a vast corpus, if digitized it would be easily manageable with present technology. Even assuming that each book requires 100 MB of memory, the total storage requirement would be 10^9 GB. With current retail prices for disk storage hovering at \$0.50 U.S. per GB in small quantity, the total cost would be under \$500 million to store every book ever written, which is much less than the original cost of purchasing, scanning or even storing one copy of each of them[1].

The question I want to address is how to use such a corpus once it is created. One of our central problems as a society is how to allow developing nations to reap the benefit of the world's technological developments. The challenge of feeding the world, providing it with fresh water and keeping it healthy depends on distribution of knowledge in usable form

[1] The typical cost to store a book in a circulating library, all charges included, is about U.S. \$6.00 per book per year. At this rate, the digital storage would be paid for in a month.

to people who need it. For example, the spread of modern farming techniques to areas of chronic famine or overpopulation would be of incalculable value. This requires much more than distributing copies of English books on agriculture to people who cannot read English. Yet the historical function of libraries has been primarily to allow selected populations to view or borrow a small set of chosen publications.

The media visionary Marshall McLuhan observed that the first use of new technology is imitation of the old. History is so replete with examples of this phenomenon that no one seems to have asked whether it is simply an observed human reaction or a fundamental necessity. It appears to be fundamental. When a new technology becomes available, it cannot displace the old unless it is able to replicate every one of the old functions. Otherwise, it will not be adopted. This creates a low threshold for acceptance. All the new technology must do is to duplicate the old, possibly faster, cheaper or with enhanced performance, and it is likely to be embraced. There is no economic imperative to have it do more than that.

Translating McLuhan's maxim into the digital library context, we can expect that the first large digital libraries will simply replicate traditional ones. That is, they will accumulate carefully selected works, index them, and deliver them, or allow them to be viewed, under controlled conditions to authorized patrons. The indexing will be excellent and the delivery remarkably rapid, but nothing really new will happen. While this is undoubtedly a desirable development, I seriously question the degree to which it will benefit mankind.

The reason is that mere access to books does not solve societal problems. The lifetime reading capacity of a human being is about 3 000 books. This assumes reading one book per week for 60 years. A very diligent reader might push the number to 10 000 books in a lifetime, which is still only 1/100 of one percent of what has been published. And how much can the reader actually assimilate from all this reading that will train him for a new job, improve his crop yields or teach him mathematics?

It is important here to distinguish reading for pleasure from reading for other purposes. The joy of reading poetry, or novels, or even well-expressed technical material, will continue unabated as it has for thousands of years. Satisfying an appetite for pleasure is not one of the principal objectives of the Universal Digital Library, however. Its greater concern must ultimately be for those who need information and knowledge for practical purposes to better themselves, and this implies a far different manner of use of its content.

The reality is that most references humans make to textual materials are not the result of large-scale reading or assimilation, as one might observe in college students reading assigned textbooks, but are in the form of directed lookup, that is, the result of search. When one has a question, one needs the answer, and often does not have sufficient time to read entire works to obtain it. The farmer who needs to rid his field of a specific pest is not interested in the history of agriculture, nor even of prescriptions for killing the multitude of insect species that have plagued farmers for centuries. He wants to know what to do today to solve his problem, and that is a search question.

If the answer is contained in one paragraph of one book written in a foreign language, the farmer is not interested in paying for a full translation of the foreign book into his own language so he can read the relevant paragraph. He just wants the answer. Unfortunately, the copyright system does not provide an efficient mechanism to allow him to pay the proportional value of the text he wants to read in comparison with the price of the whole work. Under the present mechanism, he must locate the book, which is currently not easy unless it has been digitized, then either buy or borrow the book and commission of translation of the portion he desires. It is illegal for his library to make a digital copy of the paragraph available to him without permission of the copyright owner, and the cost of even asking for such permission will often exceed the price of the work.

That method might succeed if copies of books were always readily and cheaply available. The UDL can eventually scan all public domain[1] works to solve part of the problem. Without the need to restrict distribution or account for royalties, every literary work on Earth or any portion of any such work will be free to us all. This is a major advance.

[1] "Public domain" is an often misunderstood term that means "free of copyright and available for any otherwise legal use without charge." It does not simply mean "accessible to the public".

However, because of the gradual lengthening of copyright terms around the world, particularly in the United States, which has extended the term considerably beyond what is required by international conventions, by far the majority of works so far published on Earth are still in copyright. To estimate the percentage at 90% would not be amiss. Of even greater concern is that technological materials having the most relevance are very recent and hence most likely to be in copyright.

Even copyright itself would not be such a huge problem were it not for the fact that a large percentage of works that are still in copyright are out of print. This puts the potential user in an impossible situation. Even if he is willing and able to buy the book, he literally cannot do so. The publisher is unable to supply a copy, and to duplicate a library copy would constitute copyright infringement. Even if the potential user were willing to risk an infringement suit, the UDL cannot put itself in the position of contributing to the infringement.

Prof. Raj Reddy of Carnegie Mellon University, who conceived of the UDL, has suggested that owners of out-of-print works ought to consent to an arrangement in which a user of out-of-print in-copyright (OPIC) material would pay a fee to download or print it, which would be split between the UDL and the owner. In case the user wanted a hard copy of the work, the need could be satisfied by an on-demand printing house that would sell him a single copy (produced from the digital version at the UDL). The proceeds would be divided among the owner, the UDL and the printing house.

This solution has encountered obstacles, the foremost of which I refer to as the "Titanic" problem. Numerous books on the sinking of the Titanic have appeared since 1912. Most go out of print in short order. Others, such as Walter Lord's "A Night to Remember," became best-sellers before sales diminished to a negligible level. When the movie "Titanic" was released in 1997, several of these books returned to the best-seller list. The publishers argue that if they had a revenue-sharing contract with the UDL, they would miss out on the revenue spike associated with such events.

The result is that most publishers will not agree to the use of their works in such a manner. A secondary problem is that the OPIC proposal does not address uses of the work other than manufacture and distribution of whole copies.

COPYRIGHT ISSUES

It is clear that copyright is a major impediment to the development of digital libraries. There is also considerable uncertainty in most countries as to the scope of fair use of copyrighted works and it is uncertain whether any form of digitization, even just for indexing, is permissible. Even in the United States, which has produced the most extensive set of legal determination on this issue, there is still reluctance by small organizations to test the limits of the law.

To demonstrate that the question does not admit of a simple answer, consider scanning a copyrighted work only for the purpose of developing an online keyword index of its text. The U.S. Digital Millennium Copyright Act (DMCA) provides that, under specified conditions, a "service provider shall not be liable … for infringement of copyright by reason of the provider referring or linking users to an online location containing infringing material or infringing activity, by using information location tools, including a directory, index, reference, pointer, or hypertext link..."[1]. This is the statutory justification for large-scale indexes. Even if the index points to an infringing copy of a work, the indexer is not liable unless it knows of the infringement.

One might assume that, if it is legal to create an index of a copyrighted work, it ought to be legal to distribute the index to others. The index itself is not copyrightable, and, after all, others would be privileged to generate an index themselves, so what could be improper about sharing indexes? The difficulty is that an index to a work, which typically contains the exact position of every word in the work, can be used to reconstruct the work in its entirety, and distributing the entire index, as opposed to simply allowing online queries, can be regarded as equivalent to distributing the work itself.

A digital library site has great difficulty complying with copyright law, even if it wants to. A first problem is that it is extremely hard to determine whether a work is still subject to copyright, and, if so,

[1] 17 U.S.C. §512.

who the copyright owner might be. In the U.S., diligent examination of copyright registration and renewal records is insufficient to produce an answer, even for works for which registration was sought. Since the elimination of copyright formalities by international agreement, there is no office or database to which one might refer that gives the copyright status of a work. One reason is that the ownership of copyright may have been transferred to another party. While such transfers must be in writing, there is no requirement that the Copyright Office, the repository of copyright records, need be informed of the transfer. There is thus no place one may look for a definitive determination of ownership, so a digitizer who wants to seek permission is frustrated form doing so.

Suppose one learns that a work is still in copyright, but copyright owner no longer exists or cannot be located. Is it legal to copy the work, or is it simply taking a risk that is unlikely to produce adverse consequences?

PIRACY

The book world has remained remarkably unscathed by the sort of digital piracy that has plagued the movie and record industries. There is no Napster for books. I believe that the reason is not a low level of interest in reading, but the fact that it is still painful to read from computer terminals while it is not at all difficult to watch DVDs or listen to music on digital devices.

The situation will change radically when an electronic book (eBook) artifact is available that mimics the properties of a paperback book, namely the ability to flip and bend pages, mark one's place, write notes in the margins and put it in a convenient pocket. Of course, this device will be able to store 10 000 books (by our count, requiring less than a Terabyte)—a lifetime supply for our diligent reader. Alas, McLuhan tells us that the first of these will do little more than pretend they are traditional books—at least until they win acceptance.

Once the true eBook exists, the publishing industry will fight precisely the same battle now being waged by music and movie companies: it will be impossible to put the e-genie back in the bottle and piracy will run rampant in this domain as well. In the meantime, development is proceeding rapidly on digital rights management (DRM) technology, which copyright owners expect will make piracy impossible.

I believe this to be a futile hope. Any work intended for humans must be presented in a form that the human senses can perceive. For books and videos, this means that the content must be visible. For music, it must be audible. Therefore, at the point that the digital material is transformed into analog form for presentation to the human, it can be captured. It cannot be protected from capture because the brain does not have a direct digital interface. The unencrypted signals must go through the air so the content can be acquired by the human. This is the reason that DRM is doomed to failure. Nevertheless, it will surely present barriers to the use of digital materials.

COMPULSORY LICENSING

Different solutions to the copyright problem have been proposed. Some are promising only within McLuhan's inhibitory prescription, that is, so long as the lending library model dominates. The United Kingdom has implemented a "public lending right", (PLR) which recognizes the fundamental unfairness of purchasing one copy of a book, placing it in a public library, and having large numbers of people benefit from it through borrowing without any additional payment to the copyright owner. The owner is compensated only once, at the initial purchase of the book. The UK Parliament each year allocates money to a common fund, the proceeds of which are distributed to copyright owners on a pro rata basis depending on how often each book is checked out of public libraries. The success of the scheme depends, of course, on the amount allocated. It might be *de minimis*, or it might constitute a windfall.

Regardless, the measurement is based on checking out entire books and is unable to measure partial uses, such as repeated reading of a single critical paragraph or section. However, the public lending right is increasing in popularity around the world, as it deserves to be, since it corrects a basic inequity.

The digital analog of a public lending right is difficult to formulate, since it, too, suffers from an inability to measure, or provide compensation for,

partial uses. A critical assumption underlying the PLR is that the publisher benefits from revenue for each copy of the work that is placed in a library, and that each copy can only be lent out to one person at a time. Therefore, if the demand for a book increases, more libraries will have to buy it, and the publisher will benefit directly.

This assumption does not hold for digital works. One copy in digital form can be distributed at essentially no cost to an arbitrary number of users in essentially no time. Furthermore, many users will only want to view, use or download only a single page from a lengthy work, and even if proportional compensation were provided to the publisher, the result might be unfairly low.

A different sort of solution is the compulsory license. This is a permission to reproduce copyrighted material which the owner cannot refuse, but the user must pay for. How much the user pays can be determined by a statutory formula, as in the United States, or by an independent tribunal, as in Japan.

The U.S. has an effective, but narrowly limited, compulsory licensing scheme for two kinds of works: phonorecords of copyrighted songs and redistribution of television signals over cable systems. Anyone who wishes to sell copies of a sound recording of a song may do so without prior permission by simply notifying the copyright owner and remitting a statutory royalty based on the length of the song and the number of copies produced[1].

These compulsory licenses work because the quantum of use is easily measurable—it is determined by the number of copies made and the duration of the work. However, no ancillary rights are included in the compulsory license, such as the right to make derivative works like translations. Again, it is difficult to see what the digital equivalent of such a license would encompass since there is no useful digital analog of "copy".

The Japanese compulsory license is much more extensive, but also more cumbersome. A far larger category of works is subject to compulsory licensing, but the royalty is determined by an adjudicatory body rather than statute, which introduces delay and uncertainty in the process.

[1] The royalty in the U.S. for each copy as of January 1, 2006 will be 1.75 cents per minute with a minimum payment of 9 cents per song.

Of course the difficulty with any compulsory licensing scheme is collecting revenue and managing payments. If governments were to contribute substantial amounts of money to a common fund to pay copyright owners for the use of their works, the problem would be far simpler. Few, if any, governments understand that paving the information highway is as important as maintaining the traditional one.

SCIENTIFIC JOURNALS

Possibly the most persuasive argument against copyright arises when the use of the law acts in a manner directly contrary to its stated intent. The underpinning of copyright law in the U.S. is the clause in the Constitution giving Congress the power to enact copyright legislation: "Congress shall have power … to promote the progress of science and useful arts by securing for limited times to authors and inventors the exclusive right to their respective writings and discoveries" (Pollack, 2002). The term "science," as used at the time, was not restricted to scientific activity as we now understand it, but referred to learned studies in general. The definition is thus broader than it might appear.

The Constitutional purpose behind copyright is inhibited if publishers of scientific material are able to keep the results of research out of the hands of academics by charging exorbitant (actually, prohibitive) prices for scientific journals. An extreme example is Elsevier's oft-cited journal *Brain Research*, whose one-year subscription price for 2005 was U.S. $23 483, as reported on the company's website.

Few would begrudge a publisher the opportunity to make a profit, and although the number of subscribers to *Brain Research* is undoubtedly low, the economics of academic journal publishing set it apart in that the authors, reviewers and editors receive no compensation for their work—only the publisher gets paid—and the original purpose for which the work was created, which is to achieve widespread dissemination rather than earn money, is thwarted, and even choked off, by the fee structure.

There was a time when scholars were completely dependent on print publishers to distribute their work, and many researchers remember a day when they would await anxiously the arrival in the post of the

next issue of a journal in their field. Print publication to achieve dissemination is no longer required, or even efficient. What costs $23 483 to buy on paper costs $0 to produce on the Internet, since everyone connected with the creation of a paper works for free. Publishers now typically require camera-ready copy of papers, or at least an electronic format, so even the typesetting is without charge. This means that the useful business future for expensive print journals can be measured by an hourglass, and I hope their death can be hastened by the UDL.

THE IDEA/EXPRESSION DICHOTOMY

As a general rule, subject to international complexities discussed below, copyright law does not protect ideas, facts, or processes, but only the manner in which they are expressed. For example, the U.S. Copyright Act states, "In no case does copyright protection for an original work of authorship extend to any idea, procedure, process, system, method of operation, concept, principle, or discovery, regardless of the form in which it is described, explained, illustrated, or embodied in such work"[1]. This is a wonderful exception, since what we want from copyrighted works for the purpose of practical benefit to mankind is precisely what is uncopyrightable, namely the knowledge, ideas, concepts and operational instructions they contain. The question, explored below, is the extent to which the uncopyrightability of ideas is recognized outside the U.S.

In some cases, facts or ideas admit essentially of only one possible expression. In such cases, the idea wins out over expression and the expression becomes uncopyrightable according to the doctrine of "idea/expression merger". To hold otherwise would allow the discoverer of a fact to obtain a monopoly over it for the full term of copyright. For example, suppose someone discovered that "eating peaches cures cancer" and published a pamphlet containing that phrase. All ways of expressing that discovery are essentially equivalent. In the parlance of copyright law, they are "substantially similar". If the phrase were copyrightable, then the owner might extract a royalty from anyone who published such an instruction until the expiration of copyright, in effect granting the protection afforded by a patent, but for a much longer period of time. Since this is undesirable, the idea and expression are said to merge into the idea alone, and no copyright protection is possible.

INTERNATIONAL COPYRIGHT AGREEMENTS

Copyright law is largely territorial—acts that take place in a given country are subject to the copyright law of that country and no others[2] (Pa, 2000). However, large numbers of nations have entered into a series of treaties that provide for certain minimum levels of copyright protection and accord to foreigners the same level of protection as that provided to nationals of the country, a concept referred to as "national treatment." The principal agreements are the Berne Convention, the Universal Copyright Convention (UCC), the Paris Convention for the Protection of Industrial Property, the WIPO Copyright Treaty and the Agreement on Trade-Related Aspects of Intellectual Property Rights (TRIPS). Any sort of analysis of the effect of these agreements on copyright law around the world is far beyond the scope of this article, save for two points: the treatment of ideas vs. expression and compulsory licensing.

The Paris Convention of 1883 established an international union now known as the World Intellectual Property Organization (WIPO), which currently has 169 members. The Paris Convention itself does not address copyright, but the Berne Convention, administered by WIPO and having 156 signatories, does so. The WIPO Copyright Treaty (WCT), which modifies the Berne Convention, has 56 members. The UCC has been adopted by 64 nations. TRIPS is an agreement between WIPO and the World Trade Organization (WTO), which has 148 members. China, India and the U.S. are signatories to all of these agreements except that China and India have not ratified the WCT.

The Berne Convention specifically allows nations to impose compulsory licenses on certain types of works (e.g., musical works), forbids such licenses on others (e.g. cinema) and is silent on other works

[1] 17 U.S.C. §102 (b).

[2] Exceptions are recognized when acts committed in one country result in infringement in another country.

such as literary works, including books and computer programs (where they are copyrightable).

The UCC does not address copyright in ideas, but contains several clauses intended to promote compulsory licensing. For example, Article V allows signatory countries to grant a compulsory license to make translations. This is a very important license since without it no one has an independent right to translate a work. The practical effect of keeping the right of translation with the copyright owner is to vastly restrict the spread and utility of the vast majority of works. Even if the publisher has no interest in publishing a Polish translation of a French work, a Polish citizen cannot do so without permission[1], and thus for the entire term of copyright the work will remain inaccessible to Poles.

The ability to have works translated is essential to the goals of the UDL. Unfortunately, the UCC only permits compulsory licenses—it does not mandate them, and relatively few nations have enacted a compulsory license for translation. As a general principle, where intellectual property treaties specifically allow wealthy nations to grant benefits to poor nations, they decline to do so, and the effect is a continual widening of the gap between the industrial and the developing world. It is this deliberately maintained inequality I propose to correct later in this paper.

TRIPS Article 9.2 reads, "Copyright protection shall extend to expressions and not to ideas, procedures, methods of operation or mathematical concepts as such". Article 10 provides that "Compilations of data or other material, whether in machine readable or other form, which by reason of the selection or arrangement of their contents constitute intellectual creations shall be protected as such." It then goes on to clarify that protection does not extend to the data or material itself, unless the material is separately copyrightable.

The WCT Article 2 states, "Copyright protection extends to expressions and not to ideas, procedures, methods of operation or mathematical concepts as such". Even though China and India are not signatories to the WCT, TRIPS contains an equivalent provision. As we shall see, these treaty provisions furnish a legal pathway to universal benefit from the UDL.

RESPONSE TO COPYRIGHT

The foregoing sections have been a lengthy prelude to my central proposal, which is that since copyright poses so many obstacles to the goals of the UDL, and we cannot ignore copyright, the only alternative is to circumvent it.

Various proposals have been made to modify the copyright statutes in various ways to promote assistance to developing nations, such as expanding the scope of compulsory licenses, relating license fees to a nation's per capita income, providing tax benefits to donors who dedicate their copyrights to the public domain, implementing micropayment schemes to provide compensation for partial use of a work, and the like. Such proposals, even if they found favour, which they do not, would take a very long time to enact because of the fierce debate that accompanies any modification to the economic effect of copyright. While the public lending right is gaining favour around the world, it only benefits citizens of the enacting country, and only applies to publicly-available physical copies of works.

I propose instead that we at the UDL operate completely within the existing statutes and international agreements and make full use of the exemption of facts, concepts and principles from the scope of copyright protection, as provided in international treaties.

It is established in various jurisdictions (e.g., the U.S.) that one is privileged to use a copyrighted work to extract its unprotected content[2]. This is the basis on which reverse engineering and digital indexing are permitted. This means that it is legal to have computers process works to obtain their essential information, provided that this can be done without copying their expression. However, this principle by itself accomplishes nothing, since the effort of distilling expression from a work is very substantial and, up to now, has required human labor[3].

[1] Both France and Poland are signatories to the Berne Convention and thus Poland must accord as much protection to the French author as it does to its own nationals.

[2] It is uncertain whether one may actually copy the work in order to engage in this activity, but it is certainly permissible to use an authorized copy.

[3] The resulting expression would also be separately copyrightable, if produced by a human, which only compounds the problem.

SYNTHETIC DOCUMENTS

A "synthetic document" is one that is produced by a machine based on other inputs, typically textual articles. This arena has been the subject of much research. Perhaps the simplest example of a synthetic document is an index, which is very laborious to generate by hand but can be created very quickly by a machine. Such an index is synthetic because it did not exist beforehand.

The reason indexes are easy to make automatically is that they are almost purely syntactic. No semantic understanding of the text being indexed is needed to create the index. There are some interesting syntactic challenges, such as recognizing whole phrases and idioms, and determining whether a word that occurs at the beginning of a sentence should remain capitalized in the index, but highly useful indexes can be produced even ignoring such issues. Google is perhaps the extreme example of how useful an index can be.

Abstracts and summaries are more advanced forms of synthetic documents since they require processing at the sentence, rather than the word, level. Detection of topic sentences and elimination of redundant content are required.

At the next level up is synthesis from multiple documents, in which a program ingests different texts and processes them to generate summaries. Google News is an excellent example. It scours online newspaper feeds and produces essentially a front page containing brief headlines, along with links to the source publications. It was described at a very general level in the Web publication "Digital Inspiration" on May 31, 2005: "Google News basically crawls news sites, finds 'story clusters,' ranks the sources, figures out how prominently each source is running the story, figures out whether its a big story or a little story, figures out geographic references, and builds the pages for the various geographic and language editions". Most people are surprised to learn that the process is entirely algorithmic. No human input is utilized, save for the effort of the newspaper reporters who wrote the original stories and the editors all over the world who independently decided which stories were important enough to publish.

At a greater level of sophistication a system could produce summaries of news articles without infringing copyright by extracting the essentials from a variety of sources and synthesizing an entirely new article from the given ones.

Eventually we want to reach the stage of automatic creation of encyclopedia articles, so that a user can request an analytical article of a given length on a specific topic, such as "Tsunami Preparedness in the Indian Ocean". The program would absorb a large corpus of documents, analyze them, evaluate their sources, check for inconsistencies and consensus, and prepare a critical article of the required length.

Using digital collections as fodder for synthetic document generators is not a new idea (Wactlar, 1996). However, it does not appear to have been recognized previously that synthetic documents, if they are created without copying original expression, are free of copyright restrictions. Not only can they be freely distributed, but they can be translated into an arbitrary number of languages.

Great strides have been made recently in automated translation (Kanellos, 2005). While translation of unrestricted, pre-existing text remains difficult, nearly perfect machine translations can be made in restricted domains if controls can be placed on the text to be translated at the time it is originally generated (Carbonell, 2000). The reason is that ambiguities in syntax and vocabulary can be resolved at generation time to yield fully translatable text. Therefore, an automated translation system, working in cooperation with a document synthesizer, would be able to produce accurate output in multiple languages without the need for human editing.

All of the above systems, from indexing software to full treatise generators and automated translation systems, if properly structured, avoid the complications of copyright law and permit the informational content of copyrighted works to be digested, translated and distributed worldwide without encumbrance.

CONCLUSION

I propose that the UDL undertake the scanning of all works, even those that are in copyright. Such scanning is legal for the purpose of creating finding aids, such as indexes, and for extracting informational content. Works that are in copyright cannot be pro-

vided to the public without permission, but, as we have seen, doing so is unnecessary for the works to be of significant use. It is essential in all UDL activity that copyrighted works be protected against theft, piracy and inadvertent distribution and used only for the purposes discussed above.

All digitized works can then be used as data to software that will produce synthetic documents. Because these documents will not be copyrighted, they can be translated into any language, especially by automated means, and distributed freely throughout the world. By this mechanism I suggest that the world will obtain immense benefit from the UDL, far beyond that which was originally envisioned.

References

Carbonell, J.G., Gallup, S.L., Harris, T.J., Higdon, J.W., Hill, D.A., Hudson, D.C., Nasjleti, D., Rennich, M.L., Andersen, P.M., Bauer, M.M., Busdiecker, III. R.F., Hayes, P.J., Huettner, A.K., McLaren, B.M., Nirenburg, I., Riebling, E.H., Schmandt, L.M., Sweet, J.F., Baker, K.L., Brownlow, N.D., Franz, A.M., Holm, S.E., Leavitt, J.R.R., Lonsdale, D.W., Mitamura, T., Nyberg, E.H., 2000. Integrated Authoring and Translation System. U.S. Patent 6,163,785.

Kanellos, M., 2005. Google Dominates in Machine Translation Tests. CNET News.com, August 22. http://news.com.com/2100-1038_3-5841819.html.

Pa, W.D., 2000. Twentieth Century Fox Film Corp. v. iCraveTV, 53 U.S.P.Q. 2d 1831.

Pollack, M., 2002. Brief of Malla Pollack, Amicus Curiae Supporting Petitioners. Art. I, Sec. 8, Clause 8. http://cyber.law.harvard.edu/openlaw/eldredvashcroft/supct/amici/pollack.html.

Wactlar, H.D., 1996. The next generation electronic library —Capturing the experience. *ACM Computing Surveys*, **28**(4es). http://www.acm.org/pubs/citations/journals/surveys/1996-28-4es/a114-wactlar/.

Welcome contributions from all over the world

http://www.zju.edu.cn/jzus

- The Journal aims to present the latest development and achievement in scientific research in China and overseas to the world's scientific community;
- JZUS is edited by an international board of distinguished foreign and Chinese scientists. And an internationalized standard peer review system is an essential tool for this Journal's development;
- JZUS has been accepted by CA, Ei Compendex, SA, AJ, ZM, CABI, BIOSIS (ZR), IM/MEDLINE, CSA (ASF/CE/CIS/Corr/EC/EM/ESPM/MD/MTE/O/SSS*/WR) for abstracting and indexing respectively, since started in 2000;
- JZUS will feature **Science & Engineering** subjects in Vol. A, 12 issues/year, and **Life Science & Biotechnology** subjects in Vol. B, 12 issues/year;
- JZUS has launched this new column "**Science Letters**" and warmly welcome scientists all over the world to publish their latest research notes in less than 3–4 pages. And assure them these Letters to be published in about 30 days;
- JZUS has linked its website (http://www.zju.edu.cn/jzus) to **CrossRef**: http://www.crossref.org (doi:10.1631/jzus.2005.*xxxx*); **MEDLINE**: http://www.ncbi.nlm.nih.gov/PubMed; **HighWire**: http://highwire.stanford.edu/top/journals.dtl; **Princeton University Library**: http://libweb5.princeton.edu/ejournals/.