Webology , Volume 2, Number 2, August, 2005

Resource Gleaning, From Earlier Times to the Information Age

William W. Bostock
School of Government, University of Tasmania, Hobart, Tas., Australia

Received July 12, 2005; Accepted August 24, 2005

Abstract

Inspired by the film documentary The Gleaners and I, the paper defines two senses of gleaning: (1) generally, the collection of items in small quantities, and (2) more specifically, the collection of items missed or rejected during previous harvesting. As an activity, gleaning in both senses is a neglected but essential activity in the solving of problems of lack of resource, especially now in the Information Age. A classic example of the gleaning process is provided by psychoanalysis, which gleans information which would normally be seen as trivial or unacceptable, to be used in the analysis and treatment of a personality disorder. The need to glean information to break secret codes in World War II provided the impetus to create the world's first programmable electronic computer, the Colossus. The invention of the computer has brought in the Information Age, but in so doing, it has created a need for information gleaning. The publishing of research findings on the Internet is discussed and the current status of some gleaning software is commented upon.

Keywords

Information, Gathering, Research, Codes, Publication, Software

Introduction

In 2001 the French film-maker Agnes Varda presented her documentary *The Gleaners and I*, a film giving a modern day interpretation of the subject of Jean-François Millet's painting of 1867, *The Gleaners*. This painting depicts three women collecting some remaining wheat, barley or other grain that had been left behind after harvest. The film shows that gleaning is very much an activity of modern times as well, as people go about the business of collecting and putting to use unwanted or rejected resource, and while materially poor, these modern day gleaners are rich in humour and dignity.

Gleaning is in fact an ancient activity, long predating Millet's painting. The verb *to glean* comes to English via the Old French *glener*, which itself comes from the Latin *glan(n)are*, and possibly via the Old Irish *do-glenn*, to gather. The modern English word has two meanings (1) to gather or pick up ears of corn or other produce left by reapers or harvesters and (2) to gather or collect small quantities, to scrape together (news, facts, etc.), to find out or to discover something.

In the Old Testament, the widow Ruth the Gleaner humbled herself by gleaning in the fields of a wealthy farmer named Boaz. Boaz was attracted to Ruth and married her, and later Ruth gave birth to a son, Obed. Through Obed, Boaz and Ruth became the great grandparents of King David of Egypt (Briglia, 1932). Possibly because of its Biblical references, gleaning, and the leaving of materials for gleaning, became an accepted part of European rural life throughout the Middle Ages, and the right of usufruct was established. This was the right to use and enjoy another's property on the understanding that this use would be without destroying, damaging or diminishing the property. In nineteenth century France and other countries, usufruct rights, including the right to glean, were reduced or withdrawn, provoking resistance among peasants, as depicted in Balzac's novel of 1845 *Les Paysans*, (The Peasants) (Crummy, 1999). This unrest is the background to Millet's painting The Gleaners which is the centrepiece of Varda's film. The film is an ample demonstration of the negative evaluation placed upon gleaners and the activity of gleaning, the later elevation in standing of the Biblical Ruth the Gleaner notwithstanding. The gleaners presented and interviewed in Varda's film are clearly marginal, oppressed and neglected, and probably the victims of racist and sexist discrimination as well as social and economic. Gleaning is associated with waste and trash, and in modern society, where there is over consumption and therefore waste, the gleaners may well be critical of the society in whose margins they exist, a view, which Varda seems to endorse (Radosavljevic, 2003: 7).

This is exemplified in the film by the case of a man with a degree in biology who lives in a shelter and specialises in the scientific gleaning of foodstuffs of high nutritional value. But more than this, every evening between 6 and 9 pm, he teaches his mainly black, illiterate shelter co-residents to read and write, thus providing an example of the human gleaning for recycling among those who would, in the eyes of some, not be a human resource of great value. In the words of one commentator, it is "… a truly inspiring aspect of this human being" (Radosavljevic, 2003: 7).

Jean-François Millet, *Les Glaneurs*, 1867

Gleaning in Other Accounts

Gleaning for edible substances necessary for survival is an activity from ancient times to the present, but gleaning can also be for other left over or rejected resources such as

coal, wood or other energy sources. In the ghetto at Lodz (1941-1944), children searched in waste grounds for small pieces of coal, for sale or exchange and as a store of value. These children were known as the "coal miners" (Adelson, 1989: 133-136).

Gleaning as a Scientific Method

During World War II, a team of psychoanalytically trained psychiatrists was set up under the chairmanship of Dr. Walter C. Langer, with the purpose of carrying out a psychiatric assessment of Adolf Hitler. Basing themselves on a detailed analysis of all available documents gleaned from anywhere, and interviews with persons who had had more than a passing contact with Hitler, Langer and his team produced a secret report for the Office of Strategic Services, which was finally published in 1973. The Langer Report concluded in 1943 that Hitler was in all probability a neurotic psychopath, and went on to consider eight possible scenarios for his future. With what turned out to be great prescience, it declared that the eighth, that Hitler might commit suicide, was the most plausible outcome (Langer, 1973: 211). The implications of the other scenarios, such as a mythologising Hitler killed in battle or incapacitated by schizophrenia, or of assassination, particularly by a person of Jewish background, are spelt out in some detail and would have undoubtedly influenced Allied military planning with a level of appreciation of risks and benefits that are a special product of psychiatric training.

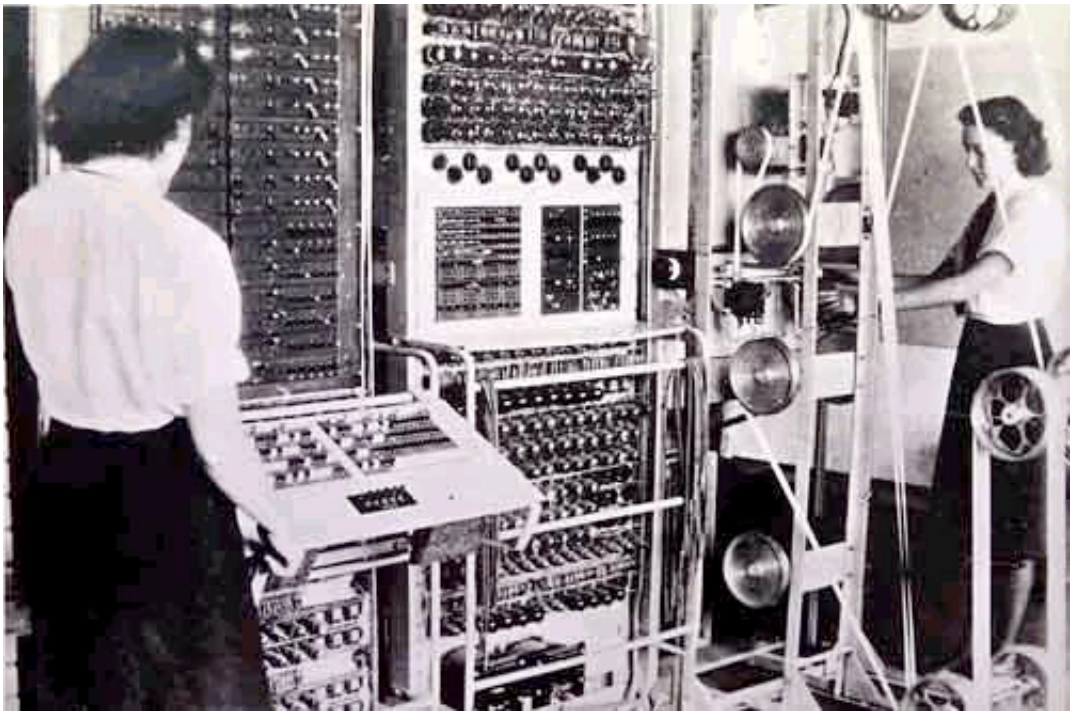
Electronic Gleaning: the Origins

The need to glean minute amounts of information of life and death significance to millions of people provided the impetus to create the world's first electronic programmable digital computer, the Colossus.

During World War II, the German military extensively used an encoding machine called *Enigma*. This machine, based on an American design, generated coded messages by substituting letters, with the serious flaw that it could not substitute for any letter the same letter. With the help of Polish code breakers, the British developed an electro-mechanical machine called the *bomba* (named after a popular brand of ice-cream). This machine was able to break the German military code and reveal many secrets.

The Germans also had a more complex cipher machine, the *Geheimschreiber* (code named "Fish" by the British), for communications of the absolute highest level such as Hitler's war directives and messages between the German Foreign Office and Germany's embassies in neutral countries. This machine used the Lorenz code and converted messages to a digital form by using rotors with changeable settings, and also had the benefit of eliminating the cyber clerk, another weakness in the encryption process. A message of 2,000 characters could generate 3,998,000 combinations, and so an electronic machine, called the *Heath Robinson*, was built, using 30 to 80 radio valves. From this was developed in 1943 the first *Colossus*, using 1,500 valves, to simulate the rotation of the Geheimeschreiber rotor wheels by counting in binary arithmetic using Boolean logic to decode the original message. Ultimately a Mark II Colossus and nine other examples were built, each using 2,500 valves and possessing an input reader that could scan 5,000 characters per second, so that when five Colossi readers were used in parallel, a reading speed of 25,000 characters per second could be

attained (Johnson, 1978: 393). At the end of World War II, some 63 million of high-grade German messages had been decrypted by the British with the aid of the Colossi (Sale, 2005: 3).



***Colossus II*: Public Record Office, London**

Colossus was not the world's first example of information gleaning by mechanical or electrical means, but it was by electronic means, though the fact was kept secret for many years after the end of World War II. The development of Colossus was closely followed by the American ENIAC, first used in late 1945, and was a much larger and faster machine using about 18,000 valves. ENIAC was used for ballistic calculations, atomic energy, thermal ignition, weather prediction, random number studies, wind tunnel design, as well as other applications where there was a need to glean information by huge calculation. Both Colossus and ENIAC were program-controlled, and from ENIAC was developed EDVAC, the world's first stored-program computer (Randell, 1980).

Gleaning Among Shredded Documents

When the German Democratic Republic collapsed in 1989, its Stasi intelligence agency set about shredding its records and stuffing them into 16,000 sacks. Most of the records were personal files that had been stored at the Magdeburg Archives, but in the haste to destroy the records, the available shredding machines could not handle the task and so the records were torn into quarters page by page and stuffed into sacks. By 2003, a team of 50 civil servants had manually reconstructed the files from some 300 sacks, in a slow process that would have taken another 450 years to complete.

To speed up the process, the Berlin Frauenhofer Institute of Production Facilities and

Construction Technology created a software program that can match fragments of paper, put them into order, and classify them by security operation. However, the implications for criminal cases including murder, treason and secret contacts with West German politicians, has meant that the application of the software has not yet been fully realized. It is likely therefore that the reconstruction of the records may be an unwanted task of gleaning (Boyes, 2003).

Gleaning From Electronic Journals

The creation, transmission and storage of knowledge is a vast human activity carried out by countless individuals, groups and institutions. At the core of this activity is traditionally the publication of research findings through the printed journal and to a lesser extent books and occasional papers. The problem in the performance of this function is that of "severe restriction" because of time and cost (Edmonds, 2000). As a result, scholars, particularly those who are younger and seeking to enter academia, can feel great frustration. Individuals can be confronted with problems of delay and even non-response as well as costs such as purchase of hard copy journals, postage, sometimes administration fees, and physical access to libraries. Moreover, minority viewpoints can be suppressed by prevailing editorial power-holders. The prestigious journals are selective, with some having a rejection rate of 90 per cent (Getz, 1997), but such is their indirect power, that there is never a shortage of submissions. The Web has provided a new location for the published products of the academic process, and in so doing, offered huge opportunity for gleaning by electronic means of research findings. A major help in advancing the acceptance of this type of publishing of research findings is the possibility of "hyperlinking", where hypertext links are provided in an article to works cited in the same article, a quantum leap forward from the traditional reference lists of hard copy articles, though the transition to clickable links is still fraught with social, commercial and legal difficulties, not to mention the conceptual one of relevance (Hitchcock, Quek, Carr, Hall, Witbrock and Tarr, 1998).

However, much of the material published on the Web is doubtful quality, reliability or authenticity, and in response to this, mechanisms for peer-review have been developed (Harnad, 1996: 7).

The creation of an association of peer-reviewed electronic journals in a specific area is a development likely to advance the reputation of quality. The *Association of Peer-Reviewed Electronic Journals in Religion* is such a body, lying down criteria for membership and thereby offering a guarantee of quality (APEJR, 2000). An interesting variation on the peer-review process is the *post-publication review* whereby articles are published as received and then voted on for publication in a more prestigious electronic archive (Nadasdy, 1997).

Recognition by an official body, which is part of government, is an unquestionably valuable indicator of quality in those countries where it is available. An example is the *Web Journal of Modern Language Linguistics* http://wjmll.ncl.ac.uk/, which is recognised by the Higher Education Funding Council of England for the purposes of recognised research output, and is therefore a valid indicator of achievement in matters of appointment, promotion and granting. Thus the gleaner of electronic academic is not

without some mechanisms of reassurance, but there remains the fear of increased vulnerability to theft of intellectual property.

Gleaning and Intellectual Property Theft

Gleaning has never been the same as stealing, and although often looked down upon socially, gleaning does not break a universal moral code, as for example stealing does. Unfortunately, the ease of gleaning with the aid of a computer has greatly increased the possibilities of intellectual property theft and plagiarism. This has created a reluctance to publish new research findings on the Web, such as those presented in a thesis or dissertation.

The electronic submission and archiving of theses and dissertations has the potential to display academic work more widely and deeply into the public sphere. In addition the development of the World Wide Web, which allows information to be shared over the Internet, the electronic thesis has enabled multi-media presentation of material as text graphics, animation and sound, in an integrated way. Until the mid 1990's, multimedia applications were uncommon due to the cost of the hardware required, but now nearly all personal computers (PC's) are capable of displaying video, as well as style sheets, linked Excel tables, animated menus, image maps, sound files and colour-coded indexes allowing information to be organized in non-linear ways. However, because of the storage demands of multimedia applications, the most effective medium is the CD-Rom, thus providing a convenient physical unit for the library storage and use of a thesis that can now moreover easily include a practical component.

There is also the possibility of hypertext links, which enormously expand the amount of information referenced in the main text. These new types of information display can be easily embodied in the work, so that interpretive capability, which functions by making connections, is greatly enhanced.

If the paperless thesis is posted on the Web, comments by readers can be received and also posted, while the potential readership expands from possibly only several people (examiners), to a readership of potentially millions, thus extending the mission of the universities in a way completely in conformity with postmodern thinking. In contrast with earlier times when traditional print theses and dissertation would average only a few request a year, some are now reported to be downloaded thousands of times. But with this exposure comes the risk of unwelcome gleaning by individuals, organizations and even governments.

In the U.K. and world wide there has been media discussion of the use of a thesis by Ibrahim al-Marashi, which was published on the Web, as the basis for a joint Intelligence Committee dossier on which British foreign policy military intervention in Iraq in 2003 was partly based, the so-called "dodgy dossier" (Wintour, 2003). The thesis is reported to have been directly copied, without acknowledgement, and including typing and grammatical errors (Jones and Williams, 2003). While this one incident does not of itself form a case against open access to academic material, it raises the question whether the author of a thesis would welcome this kind of attention, including possible plagiarism.

Some worry has been expressed at the possibility of plagiarizing charts and graphics; as a doctoral candidate at the University of Texas at Austin wrote "unless students are careful to restrict access to their work, they may quickly find their charts, graphs and maps appearing all over the web" (Perramond, 1989).

Another concern is whether graduate students would be confident that they have sufficient intellectual maturity to have their work exposed to gleaners in the way that electronic publication would allow, particularly if they appreciate that they are likely to mature in their thinking at a later date. In the arts and social sciences this may be a problem: for example, there is a distinction between "early Picasso" and "later Picasso", and most other great artists which is widely acknowledged.

While many established academics may privately (or even publicly) admit that their thesis may have lacked maturity, the level of exposure created by publication on the Web may be a discouragement. One scholar, however argues that Web publication may be a quality controlling mechanism

> 'One might conclude that research that does not measure up to peer-reviewed standards can slide by, as long as the results or that research are not exposed to wide audience via electronic distribution' (Lang, 2002: 687).

Against the attraction of gleaning by experts and genuinely interested persons through exposure of the Web, must be balanced a sense of apprehension at undesired gleaning, which may affect supervisors, examiners, schools and universities as well as candidates, and thus the possibility of access denial has been an important component in the protocols of electronic dissertations and theses that are emerging.

Open Access

A major problem confronting electronic gleaners is the growing number of multinational publishers who are now charging substantial sums of money for access to their electronic journals. These include Elsevier, Springer, Thomson, Kluwer and Taylor and Francis, who are publishing research provided to them generally free of charge, and thus requiring individuals and institutions to buy it. The Budapest Open Access Initiative arose from a meeting convened in Budapest by the Open Society Institute (OSI) in 2001. The purpose of the meeting was to accelerate progress in the international effort to make research articles in all academic fields freely available on the Internet. The participants represented many academic disciplines and came from many nations. They explored how the separate initiatives could work together to achieve broader, deeper, and faster success and the most effective and affordable strategies for serving the interests of research, researchers, and the institutions and societies that support research. Finally, they explored how OSI and other foundations could use their resources most productively to aid the transition to open access and to make open-access publishing economically self-sustaining. The result is the Budapest Open Access Initiative, (http://www.soros.org/openaccess/), a statement of principle, strategy, and commitment.

Gleaning Software Initiatives

As well as the well-known search engines, specialized programs have been developed to deal with the complex problems encountered when mining the Web (Kosala and Blockeel, 2002). Kieras, Wood, Abotel and Hornof have produced a computer-based tool for the rapid evaluation of the GOMS (Goals, Operators, Methods and Selection Rules) model of human task performance. Engineering models of human performance permit some aspects of usability of interface designs to be predicted from an analysis of the task, and thus can replace to some extent expensive user testing data. The computer-based tool, GLEAN, generates quantitative predictions from a supplied GOMS model and a set of benchmark tasks. GLEAN can reproduce the results of a case study of GOMS model application with considerable time savings over both manual modelling as well as empirical testing (Kieras et al., 1995).

Another system is called *Glean*, and has been developed by Chandrasekar and Srinivas. This works on the concept that a body of text in natural language will contain latent information such as patterns of language use and syntax, which can be used to enhance information retrieval through a standard Web search engine or Information Retrieval system. *Glean* uses a tool called a supertagger to induce patterns of information to appear, by identifying relevant information and filtering out irrelevant material. Using as an example the gleaning of information about official appointments to positions, Chandrasekar and Srinivas were able to filter out upwards of 80 per cent of irrelevant documents (Chandrasekar and Srinivas, 1997).

Another type of program is Information Extraction (IE). This deals with the problem of information overload on the Internet, which is swamping normal research and scholarship activity. What have been developed are applications that automatically merge information from many Web sites. But a problem is that data needed by integration applications is usually designed for human consumption, and rarely available in an agent-friendly form such as XML. Integration applications usually have a wrapper that translates a source's original format into a format suitable for integration. Typically, a wrapper will sift through a source's HTML, retaining the useful text, while discarding advertisements, formatting tags, and other unwanted material. While wrappers are relatively simple programs, writing them by hand is tediously labour-intensive and liable to contain errors. Kushmerick has developed a set of wrapper induction algorithms for automatically learning wrappers from examples. Experiments have shown that a large class of Web sources can be automatically wrapped using relatively few training examples. While the techniques work well with rigidly structured HTML, recent experiments show that the techniques can be extended to handle natural language text such as e-mail and printed articles. In addressing the problem of wrapper maintenance, Kushmerick has noted that Web sources frequently "remodel" their user interfaces, which usually renders wrappers ineffective. As a first step toward "adaptive" wrapper induction, he has developed an algorithm for determining whether a wrapper is operating correctly (Kushmerick, 2002).

Conclusion

Gleaning for resources such as food or energy is an ancient and honourable activity,

often necessitated by the needs of survival in extremely adverse conditions of the physical, social or political environment. In addition, gleaning for information is a valuable and sometimes essential research methodology, in fields as diverse as psychoanalysis and cryptology: in fact, there is in all branches of scientific enquiry a valid application for gleaning.

The need to glean information from secret codes in World War II provided the impetus to develop a new kind of gleaning machine, which became the world's first programmable electronic computer. From those origins have developed today's computers that are used in almost every type of human activity, including, to take just one example, the reconstruction of deliberately shredded documents. The development of the computer and the arrival of the Information Age and has also seen the creation of the Web, with huge possibilities of publishing research findings such as those in electronic theses, dissertations, reports, articles or books, which are instantly available world-wide, with corresponding risks of theft of intellectual property. There is also the new challenge to develop gleaning software that can make manageable the task of gleaning in the information age.

References

- Adelson, Alan and Robert Lapides, (1989). (Comps. and eds.). *Lodz Ghetto, Inside a Community Under Siege*. New York: Viking.
- APEJR (Association of Peer-Reviewed Electronic Journals in Religion), July 2, 2000. http://rosetta.reltech.org/apejr/apejr.html
- Boyes, Roger, (2003). Stasi puzzle-solver puts fear into many. *The Australian*, June 6, 9.
- Briglia, Margarethe Groer, (1932). Ruth the Gleaner. *Christian Science Sentinel*, July 16, 906.
- Chandrasekar, R. and B. Srinivas, (1997). Gleaning information from the web: Using syntax to filter out irrelevant information. In *AAAI Spring Symposium on Natural Language Processing for the World Wide Web*. Stanford University, (March 1997). Retrieved July 14, 2005, from http://citeseer.ist.psu.edu/srinivas97gleaning.html
- Crummy, Ione, (1999). The Subversion of Gleaning in Balzac's *Les Paysans* and in Millet's *Les Glaneuses*. *Neohelicon*, XXVI, 1, 9-18.
- Edmonds, Bruce, (2000). A Proposal for the Establishment of Review Boards. *Journal of Electronic Publishing*, v5, i4, pp10. http://www.press.umich.edu/jep/05-04/edmonds.html
- Getz, Malcolm, (1997). An Economic Perspective on E-Publishing in Academia. *Journal of Electronic Publishing*, v3, i1, 25. Retrieved July 14, 2005, from http://www.press.umich.edu/jep/archive/getz.html
- Harnad, Stevan, (1996). Implementing peer review on the Net: Scientific quality control in scholarly electronic journals. In Peek, R. and Newly, G. (eds.). *Scholarly Publication: The Electronic Frontier*. Cambridge, MA: MIT Press, pp103-108. Retrieved July 14, 2005, from ftp://ftp.princeton.edu/pub/harnad/Harnad/HTML/harnad96.peer.review.html
- Hitchcock, Steve, Quek, Freddie, Carr, Leslie, Hall, Wendy, Witbrock, Andrew

and Tarr, Ian, (1996). Towards Universal Linking for Electronic Journals. *Serials Review*, v24, n1, pp21-33. Retrieved July 14, 2005, from http://www.mmrg.ecs.soton.ac.uk/publications/archive/hitchcock1998a/

- Johnson, Brian, (1978). *The Secret War*. London: Arrow Books.
- Jones, Gary and Alexandra Williams, (2003). Real Authors of Iraq Dossier Blast Blair, *Mirror.co.uk*. Retrieved July 14, 2005, from http://www.mirror.co.uk/news/allnews/page.cfm?objectid=12620001&method=full&
- Kieras, David E., Scott D. Wood, Kasem Abotel and Anthony Hornof, (1995). GLEAN: a computer-based tool for rapid GOMS model usability evaluation of user interface designs. *Symposium on User Interface Software and Technology Archive*. Pittsburgh, Pennsylvania: Proceedings of the 8th annual ACM symposium on User interface and software technology. Retrieved July 27, 2005, from http://portal.acm.org/citation.cfm?id=215585.215700
- Kosala, Raymond and Hendrik Blockeel, (2000). Web Mining Research: A Survey. *SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on Knowledge Discovery & Data Mining*, Palo Alto: ACM, 2, 1, 1-15. Retrieved July 27, 2005, from http://citeseer.ist.psu.edu/kosala00web.html
- Kushmerick, Nicholas, (2002). Gleaning answers from the Web. *Proc. AAAI Spring Symposium on Mining Answers from Texts and Knowledge Bases*, 43-45. Retrieved July 27, 2005, from http://www.cs.ucd.ie/staff/nick/home/research/download/kushmerick-maftkb-ss02.pdf
- Lang, Susan, (2002). Electronic Dissertations: Preparing students for our past or their future? *College English*, 64, 6, 680-695.
- Langer, Walter Charles, (1973). *The mind of Adolf Hitler / the secret wartime report by Walter C. Langer*. (Foreword by William L. Langer, Afterword by Robert G. L. Waite). London: Secker & Warburg.
- Nadasdy, Zoltan (1997). A Truly All-Electronic Journal: Let Democracy Replace Peer Review. *Journal of Electronic Publishing*, v3, n1. Retrieved July 27, 2005, from http://www.press.umich.edu/jep/03-01/EJCBS.html
- Perramond, Eric P., (1998). Letter to the Editor. *The Chronicle of Higher Education* (27 March 1998). Retrieved July 24, 2003, from http://chronicle.com/colloquy/98/thesis/07.htm
- Radosavljevic, Dunja, (2003). Agnes Varda's 1'Ecriture Feminine (film critique of The Gleaners and I). Conference, *Critical Themes in Media Studies*. Media Studies Department: New School University. Retrieved July 27, 2005, from http://beard.dialnsa.edu/~treis/perspectives.html
- Randell, B., (1980). The Colossus. In N. Metropolis, J. Howlett and GC Rota, (Eds.). *A History of Computing in the Twentieth Century*. London: Academic Press, 47-92.
- Sale, Tony, (2005). *The Colossus, its purpose and operation*. Retrieved July 27, 2005, from http://www.codesandciphers.org.uk/lorenz/colossus.htm
- Wintour, Patrick, (2003). MPs Call Campbell over Iraq Dossier. *Guardian Unlimited*. Retrieved 2 June 2003, from http://politics.guardian.co.uk/iraq/story/0,12956,981374,00.html

Bibliographic information of this paper for citing:

Bostock, W. W. (2005). "Resource Gleaning, From Earlier Times to the Information Age." *Webology*, **2** (2), Article 14. Available at: http://www.webology.ir/2005/v2n2/a14.html

---
