Information Waste, the Environment and Human Action: Concepts and Research

Fons Wijnhoven, Pim Dietz, and Chintan Amrit

University of Twente, AE Enschede, Netherlands
fons.wijnhoven@utwente.nl

Abstract. Information technology is powered by electricity. Although its impact on Green House Gasses (GHG) is still rather limited, the next decade will show an explosion of its impact because technological innovations on data communication, information retrieval and datacenter operation will not compensate the increased need for energy of information technology. This paper approaches the problem not from a technical perspective, but from the perspective of information value and the opportunities to detect and remove information waste. For this we identify several indicators of information waste and we then propose some key ideas for researching the topic.

Keywords: information value, information waste, file retention, web site quality.

1 Introduction

This essay is part of a more general interest in e-waste: the consequences of the information industry on hazardous waste. Ruth [20] states (p. 74) that "Seventy percent of all hazardous waste is e-waste, which is bulky, complicated to recycle, and sometimes contains unsafe levels of heavy metal and other dangerous chemicals." E-waste may have several causes:

1. PCs, monitors, workstations and other hardware: According to the US agency for the environmental protection (EPA): "…over 25 billion computers, televisions, cell phones, printers, gaming systems, and other devices have been sold since 1980, generating 2 million tons of unwanted electronic devices in 2005 alone, with only 15 to 20 percent being recycled" [20].
2. Software: Software can be designed and developed to improve the efficiency of energy consumption of electronic devices and computers. Microsoft claimed that its Windows Vista product saves roughly $50 per year in electricity costs per PC and thus also causes reduced GHG emissions. McAfee estimates the number of spam messages in 2008 to be about 62 trillion, which causes 33 billion KWh of unnecessary use (which is equivalent to the annual use of 2.4 million homes in the US).

M.D. Hercheui et al. (Eds.): HCC10 2012, IFIP AICT 386, pp. 134–142, 2012.

3. Data centers and servers: In 2005, the power and cooling costs of servers worldwide is estimated to be US$26 billion. Greenpeace[1] reports that many of the large data centers in the USA use electricity produced by coal, i.e. the largest GHG emission creator during electricity production.

This paper focuses on a specific under-researched aspect of e-waste, namely, information waste. Information waste are data which are unnecessary (e.g. redundant) and unusable (e.g. not understandable) and which are the consequence of human limitations of knowing which data are of no use and could thus be removed or stored on a non-direct access medium. Detecting information waste will help in reducing the energy needs of information technology and related GHG emissions. Information waste can still be regarded as a minor influencer of e-energy waste, but in the next decades we may be confronted with an explosion of it. For example, IDC expects the digital universe to grow from 0.8 ZB (one ZB is 1 trillion gigabyte) in 2009 to 35 ZB in 2020; which is factor of 44 in 10 years [25]. Thus research in this area is needed in order to be prepared for this future. It is likely that the percentage of the total amount of information that is considered as waste will grow substantially as a consequence of the increased complexity of finding information in larger databases and in the Internet. Thus a major question is how we can detect information waste?

2 The Concept of Information Waste

Information is meaningful data or meaningful representations [5], [32]. Information is a key resource for organizations, and information technology is able to hugely reduce information collection, storage, manipulation, and distribution costs. In fact, the marginal reproduction and distribution costs of digital information are nearly zero [23]. The real costs are the creation of the first copy. On a world scale, though, the energy costs are very substantial, and often information technology is run on not-green electricity[2]. With more than 200 million internet searches estimated globally daily, the electricity consumption and greenhouse gas emissions caused by computers and the internet is provoking concern. A recent report by Gartner said that the global IT industry generated as much greenhouse gas as the world's airlines - about 2% of global CO2 emissions [6]. Data centers are among the most energy-intensive facilities imaginable. Servers storing billions of web pages require power. Mobile devices and smartphones also consume internet resources and substantial energy for data communication. We are not saying here that any internet use or information service use (for business or personal needs) should be avoided, but we say that it has "carbon costs." We have to be aware of these facts, and next start using information technology most intelligently. Information services should be there to help on this matter, but we lack

[1] http://www.greenpeace.org/international/en/publications/reports/How-dirty-is-your-data/

[2] http://www.greenpeace.org/international/Global/international/publications/climate/2011/Cool%20IT/dirty-data-report-greenpeace.pdf

tools for detecting information waste and thus remove it. Thus this paper poses a key question: How can we detect information waste?

This description of information waste indicates that information waste can be found on different media (on the disks in proprietary environments of persons and organizations and on the internet) and information waste can have information use and knowledge dimensions. This implies that at least four areas of information waste can be identified, as presented in Table 1, which we discuss in the following subsections.

Table 1. Types of information waste

		Information waste media	
		Proprietary disks	**Web**
Information waste dimensions	Information use indicators	Section 3.1	Section 3.3
	Knowledge value indicators	Section 3.2	Section 3.3

3 Identifying Information Waste

3.1 Use Indicators of Information Waste in Existing File Retention Methods

Determining the value, or the lack of value of information, i.e.: information waste, is complex. We can, for instance, easily calculate the number of data available on a hard disk by looking how many kilo-, mega- or gigabytes are occupied by our documents. But this does not say much about its value, as sometimes less is better. Most current information waste research is file retention research and focusses on the analysis of statistical patterns of files.

The key assumption of file retention research is that throughout its lifecycle, the value of a file in general grows after the first stage and declines in the final stage [27]. In the final stage, the intensity of usage mostly decreases and the accessibility of the files becomes less important. But, not all types of files have the same value and the file value may evolve differently depending on the file type. Consequently, one of the most important functions of a file valuation method is the ability to differentiate files by its value and non-value in an unbiased manner so that decisions can be made on the appropriate storage medium or possible deletion of these files [2]. Hence, what is required is a method to relatively easily measure the use value of files by which a file retention (or deletion) policy can be determined. We found ten data retention policy formation methods in the literature. Table 2 gives an overview of these methods.

A number of criteria for a file retention policy method are present in the literature:

1. The retention policy determination method has to function with little to no human intervention [2], [28]. The execution of file valuation as a manual rating of individual is mostly too costly. A simple directory can easily contain 6,000 files; evaluating them piece for piece will take many hours if not days.

2. The method should be based on the subjective use value of files over time in their different life stages [2], [28]. It is obvious that value is a subjective and often individual characteristic.
3. The method has to use multiple file attributes for the valuation process [28]. One file attribute will not be able to cover all value determining variables.

Table 2. File Policy Retention Determination Methods

Author	Goal of data retention policy	Important file attributes
[2]	Capture the changing file value throughout the lifecycle and present value differences of files	Frequency of use; Recency of use
[28]	Determine the probability of future use of files for deciding on the most cost-effective storage medium	Time since last access; Age of file; Number of access; File type
[1]	Lay out storage system mechanisms that can ensure high performance and availability	Frequency of use
[30]	Optimize storage allocation based on policies	Frequency of use; File type
[14]	Classify automatically the properties of files to predict their value	Frequency of use; File type; Access mode
[34]	Select files that can be compressed to reduce the rate of storage consumption	Directory; File name; User; Application
[26]	Optimize storage in a hierarchal storage management (HSM) solution	Least recently used
[7]	Reduce storage consumption on primary storage location	Time since last Access
[22]	Design a cost efficient data placement plan while allowing efficient access to all important data	Metadata; User input; Policies
[10]	Determine file value based on supply and demand	Frequency of use (by different users)

All the file retention policy determination methods of table 2 can be automated, and thus fulfill the first criterion. They all classify files by file attributes in order to make retention decisions. In some way these methods must be able to represent file value (criterion 2) and some combination of these file attributes must be able to identify waste. File value, however, is a subjective dimension and consequently must be measured and cannot be derived from file behavior alone.

3.2 Knowledge Value

Five paradigms to the knowledge value have been codified by epistemological (i.e., knowledge theory) traditions [3]. These paradigms are:

- The empirical paradigm based on John Locke (1632-1704) [15], [29] evaluates information value by its correctness in representing facts and events in reality.
- The rationalist paradigm founded by Gottfried Wilhelm Leibniz (1646-1716) [9], [13] evaluates information by its opportunity to causally explain, predict and reason about problems and reality.
- The transcendental idealist paradigm, founded by Immanuel Kant (1724-1804) [8] evaluates information by both empiricist and rationalist criteria, but on top of that it analyzes the key a priori of the views taken and from there aims at further integrating different perspectives in a larger coherent view of a subject.
- The Hegelian paradigm developed by Georg Wilhelm Friedrich Hegel (1770-1831) [18], [24] evaluates information by its historical context and sees information as representation of conflicting interests that can be synthesized by dialect logic. As such Hegelian dialects gives concepts for interpreting human behavior and critically looking at the status quo, and as such is a foundation for interpretive [11] and critical [31] explanatory insights.
- The Lockean, Leibnizian, and Kantian paradigms of knowledge all aim at finding an ultimate truth. The Hegelian approach regards truth as part of historical and social reality, and as arguments in favor of certain ideals. The pragmatist paradigm, as described by Churchman [3] on the basis of Edgar Singer's (1873-1954) work, in contrast proposes that the continuous search for new and improved insights is important, but only valuable as far as it results in human progress, which implies the practical solving of human problems. For the measurement of the pragmatic value of information, Sajko et al. [21] developed an information value questionnaire (IVQ) that allows information workers to value the information they use. The IVQ has five dimensions (1) Files Lost, (2) Costs of File (Re)building; (3) Market Value; (4) Legislative, and (5) Time as an indicator of obsolescence. The "Lost" dimension measures the impact of information loss on the business operations. This can be anything from "nothing special" to "making wrong decisions with major consequences". "(Re)building" measures the cost of replacing the lost information (from "negligibly small" to "intolerably high costs"). "Market value" measures the consequences if competitors obtain the information (from "nothing" to "competitor gets competitive advantage"). "Legislative" identifies the obligation to keep the information and the legal consequences if the information is lost (from "no obligation" to "keeping information is obligatory and sanctions are strict"). The "Time" dimension measures the rate at which the information depreciates in value (from "very quickly" to "does not depreciate at all").

These approaches to knowledge give different indications to information waste, as summarized in Table 3.

3.3 Web Information Waste

For internet information, a number of behavioral indicators can be found and several scales for the knowledge value of sites exist. Several web analytics companies (like Google, Alexa and URLSpy) deliver behavioral data on the intensity of use of sites.

Alexa.com also publishes a top million list every day. These web data collectors produce the following behavioral site attributes as possible waste indicators:

Table 3. Knowledge value criteria for information waste

Paradigm	Information waste criteria
Lockean	No correspondence with reality; incorrect representations; not interpretable in natural language
Leibnizian	Inconsistency, wrong or obsolete parameters and fomulas, over-complex models
Kantian	Statements or content not related to an ontology.
Hegelian	Information serving no one's interest
Singerian	Irrelevant and un-usable information

- Access speed. More access speed has been indicated as poor maintenance or less professional support to the sites quality [4], [16], [33].
- One of the most important behavioral metric is the number of incoming links. If a website has a lot of incoming links, it is expected to contain good information.
- The number of broken links on a website is an indicator of maintenance problems.
- Currency can be easily measured by the last update or modification time of a site.
- Frequency of Access can be measured by the number of unique (monthly) visitors to a site. If a site has a lot of visitors it most likely is valuable information.
- Time on a site. If a user stays at a site for a long time, it most likely is good information. Precautions need to be taken with this metric because if a user keeps his browser open at a certain page while he is away, it will give a false positive.
- Bounce percentage. Bounce percentage gives the percentage of unique users which visited only one page on a certain website. Therefore this might give an indication of poor information quality.

For the knowledge value of sites, a lot of research has already been done in the area of website quality [4], [16], [33]. The research realized various metrics, both objective and subjective, to classify the value of websites. Although these metrics are relatively old for the fast changing internet, they are still used in relative new researches, for example [12], [17], [18]. At the time these studies were performed most of the current content of websites was already present. The following scales for information value have been developed:

- Content quality and correctness [4], [16] determines information on a site is correct. This is necessary to determine the information quality of a website since a low content quality means a low information quality and vice versa.
- Information relevancy [33] determines whether a site delivers relevant or irrelevant information.
- Information comprehensiveness [4], [33] indicates the completeness and understandability if a site's content.

4 Researching Information Waste

Returning to our question how information waste can be detected, we have stated that the individual rating of files or sites is too laborious for estimating the knowledge value of the content of these media. Consequently, an measurement of value on basis of behavioral indicators using analytics or file system tools can be tried, because it is much more efficient, but it is unclear until now how reliably behavioral indicators can estimate (subjective) knowledge value. If in ad random selected files and sites a high correlation can be found between some behavioral indicators and knowledge value of files, these behavioral indicators can be used as proxies for information waste.

For Internet information waste estimation, the following set of assumptions and hypotheses can be used for further research:

- Assumption 1: The higher the processing costs of servers, the higher the server's GHG footprint.
- Assumption 2: The higher the search and access costs of information, the higher the Internet users GHG footprint.
- Hypothesis 1: The higher the information waste (=% of unnecessary data on the Internet), the higher unnecessary amount of the processing costs of servers.
- Hypothesis 2: The higher the information waste, the higher the avoidable search and access costs of useful information.
- Hypothesis 3: The availability of an effective information detector will result in a reduction of information waste by increased information waste awareness of information service customers.

Figure 1 summarizes this in a causal research model.

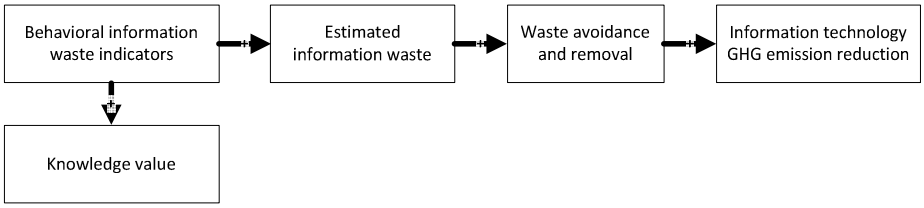

Fig. 1. An information waste design science research model

References

1. Bhagwan, R., Douglis, F., Hildrum, K., Kephart, J.O., Walsh, W.E.: Time Varying Management of Data Storage. In: Workshop on Hot Topics in System Dependability, Yokohama, pp. 222–232 (2005)
2. Chen, Y.: Information Valuation for Information Lifecycle Management. In: Proceedings of Second International Conference on Automatic Computing, Washinton, pp. 135–146 (2005)
3. Churchman, C.W.: The design of inquiring systems: basic concepts of systems and organization. Basic Books, New York (1971)

4. Eppler, M., Muenzenmayer, P.: Measuring information quality in the web context: a survey of state-of-the-art instruments and an application methodology. In: Proceedings of the Seventh International Conference on Information Quality (2002), http://mitiq.mit.edu/iciq/iqpapers.aspx?iciqyear=2002
5. Floridi, L.: Is semantic information meaningful data? Philosophy and Phenomenological Research 70, 351–370 (2005)
6. Gartner: Gartner Says Data Centres Account for 23 Per Cent of Global ICT CO2 Emissions (2007)
7. Gibson, T., Miller, E.: An Improved Long-Term File-Usage Prediction Algorithm. In: Annual International Conference on Computer Measurement and Performance (CMG 1999), Reno, NV, pp. 639–648 (1999)
8. Hartnack, J.: Kant's theory of knowledge. Harcourt, Brace & World, New York (1967)
9. Huenemann, C.: Understanding rationalism. Acumen, Chesham (2008)
10. Jin, H., Xiong, M., Wu, S.: Information value evaluation model for ILM. In: 9th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, SNPD. IEEE (2008)
11. Klein, H.K., Myers, M.D.: A set of principles for conducting and evaluating interpretive field studies in information systems. MIS Quarterly 23, 67–94 (1999)
12. Kuo, Y.F., Wu, C.M., Deng, W.J.: The relationships among service quality, perceived value, customer satisfaction, and post-purchase intention in mobile value-added services. Computers in Human Behavior 25, 887–896 (2009)
13. Look, B.C.: Gottfried Wilhelm Leibniz. Stanford Encyclopedia of Philosophy (2007), http://plato.stanford.edu/archives/spr2009/entries/leibniz/
14. Mesnier, M., Thereska, E., Ganger, G.R., Ellard, D.: File Classification in Self-* Storage Systems. In: Proceedings of the First International Conference on Autonomic Computing. IEEE Computer Society Press (2004)
15. Meyers, R.G.: Understanding empiricism. Acumen Publishing, Chesham (2006)
16. Palmer, J.W.: Web site usability, design, and performance metrics. Information Systems Research 13, 151–167 (2002)
17. Petter, S., Delone, W., Mclean, E.: Measuring information systems success: models, dimensions, measures, and interrelationships. European Journal of Information Systems 17, 236–263 (2008)
18. Popovič, A., Coelho, P., Jaklič, J.: The impact of business intelligence system maturity on information quality. Information Research 14, 4 (2010)
19. Redding, P.: Georg Wilhelm Friedrich Hegel. The Stanford Encyclopedia of Philosophy (1997), http://plato.stanford.edu/archives/sum2002/entries/hegel
20. Ruth, S.: Green it more than a three percent solution? IEEE Internet Computing 13, 74–78 (2009)
21. Sajko, M., Rabuzin, K., Baca, M.: How to calculate information value for effective security risk assessment. Journal of Information and Organizational Sciences 30, 263–278 (2006)
22. Shah, G., Voruganti, K., Shivam, P., Alvarez, M.: ACE: Classification for Information Lifecycle Management. Computer Science IBM Research Report, RJ10372, (A0602-044) (2006)
23. Shapiro, C., Varian, H.: Information rules: a strategic guide to the network economy. Harvard Business School Press, Boston (1999)
24. Sinnerbrink, R.: Understanding Hegelianism. Acumen Pub. Ltd. (2007)

25. Slawsky, D.: Teaching digital asset management in a higher education setting. Journal of Digital Asset Management 6, 349–356 (2010)
26. Strange, S.: Analysis of Long-Term UNIX File Access Patterns for Application to Automatic File Migration Strategies. University of California, Berkeley (1992)
27. Tallon, P.P., Scannell, R.: Information Lifecycle Management. Communications of the ACM 50, 65–70 (2007)
28. Turczyk, L., Frei, C., Liebau, N., Steinmetz, R.: Eine Methode zur Wertzuweisung von Dateien in ILM. In: Bichler, M., et al. (eds.) Multikonferenz Wirtschaftsinformatik. Gito Verlag, Berlin (2008)
29. Uzgalis, W.: John Locke. The Stanford Encyclopedia of Philosophy. Stanford University, Stanford (2010), http://plato.stanford.edu/archives/fall2008/entries/locke/
30. Verma, A., Pease, D., Sharma, U., Kaplan, M., Rubas, J., Jain, R., Devarakonda, M., Beigi, M.: An Architecture for Lifecycle Management in Very Large File Systems. In: 22nd IEEE / 13th NASA Goddard Conference on Mass Storage Systems and Technologies, pp. 160–168 (2005)
31. Walsham, G., Sahay, S.: GIS for District-Level Administration in India: Problems and Opportunities. MIS Quarterly 23, 39–66 (1999)
32. Wijnhoven, F.: Information services design: A design science approach for sustainable knowledge. Routledge, New York (2012)
33. Yang, Z., Cai, S., Zhou, Z., Zhou, N.: Development and validation of an instrument to measure user perceived service quality of information presenting web portals. Information & Management 42, 575–589 (2005)
34. Zadok, E., Osborn, J., Shater, A., Wright, C., Muniswamy-Reddy, K., Nieh, J.: Reducing Storage Management Costs via Informed User-Based Policies. In: IEEE Conference on Mass Storage Systems and Technologies, Maryland, pp. 101–105 (2004)