Editor: Ronald D. Williams, University of Virginia, Thornton Hall/Electrical Engineering, Charlottesville, VA 22903;
phone (804) 924-7960; fax (804) 924-8818; Internet, rdw@virginia.edu

Customizing information:
Part 1, Getting *what* we need, *when* we need it

Dan Berleant and Hal Berghel, University of Arkansas

"Dawning upon us . . . is . . . a common world brain."
— from an essay by H.G. Wells, 1938

Although he anticipated universal access to the entire body of human knowledge, even H.G. Wells might have been surprised at the sheer volume of information currently available. As we move further into the information age, it is becoming ever more apparent that society as a whole, and information and computing specialists as its agents, will have to confront the general problem of information overload. The rising flood of information will soon compel us to use techniques and resources aimed at maximizing our information-handling efficiency. Storing and retrieving digital information according to consumer requirements is only part of the equation; information must also be presented in a form suited to the consumer's needs at the time of consumption. We call this *information customization* and characterize it as the transformation of information into its most appropriate form. Thus, customization makes existing information more useful.

The information pipeline. Envisioning the information pipeline can help us understand the importance of customizing digital information (see Figure 1). The life cycle of information artifacts (and of manufactured artifacts as well) progresses in stages. Customization may occur before the last stage, which is use. Information may be used to produce another artifact, leading to another trip through the pipeline. For example, an e-mail message might be used to help produce an article or a program.

Information must be presented in a form suited to the consumer's needs.

Customization is becoming increasingly important as information artifacts flood society at an ever-increasing rate. Information customization, enabled by modern information technology's efficiency in production, distribution, and use, is poised for recognition as a critical field. The current state of information distribution — and especially its limitations — illustrates the need to customize information.

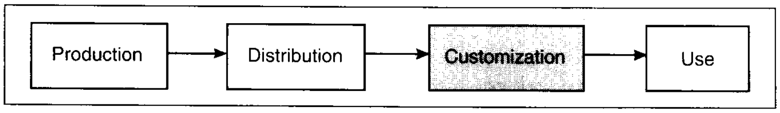

Figure 1. Customization will become an increasingly important stage in the information pipeline.

Information distribution. In the early days of electronic information distribution, information consumers interfaced electronically with information sources such as bulletin boards, networks, servers, and so forth, downloading information with relatively primitive tools like ftp. As the amount of available information mushroomed, it became necessary to retrieve information more selectively. Consequently, rudimentary Internet navigation and browsing tools (such as Gopher, Archie, Veronica, and WAIS) were implemented. These are giving way to more sophisticated hyperlink client-server software (for example, Cello and Mosaic for accessing the World-Wide Web). The age of interactive access to worldwide repositories of on-line, multimedia information is upon us.[1]

Still, information overload taxes the capabilities of even the most advanced navigation and browsing tools. New techniques under development are intended to electronically filter the flow of documents off the networks,[2] thereby fine-tuning the information distribution process. The goal is to attract the most useful documents and reject the rest. Researchers are focusing on advanced methods of attracting information using such filtering and retrieval techniques as keyword vector comparisons, latent semantic indexing, and so forth.[2] Thus, traditional information retrieval is being joined by a new and closely related discipline: information filtering.[3]

Information retrieval and filtering help the end user obtain the right documents from a larger body of available documents, but they do not solve the problem of getting the right information from *within* a given document or other information artifact. This is where information customization comes in.

Information customization. Information distribution is inherently limited. Acquiring the right document or other information artifact at the right time is certainly beneficial, but mere acquisition is not enough. An information artifact contains a mixture of information of greater or lesser value, depending on the information consumer's point of view, and manually locating the most valuable information within an information artifact is tedious and error prone.

Information customization avoids the unwanted imposition of predefined structure.

Information is customized when transformed into an extract or other form that gives a consumer what is needed at a given moment. The process may include editing, establishment of navigation links, annotation, use of browsing software, structure or keyword analysis, information visualization or animation, or other means. Information can be customized by hand, but the evolution of information technology calls for partial or even total automation.

The concerns of information customization overlap those of several other fields in which similarities and differences are evident. We have already shown that information retrieval and filtering differ from information customization in typically taking a set or stream of documents as input. Information customization takes as input a single electronic document or other information artifact. Moreover, when information is customized, we transform the document into a new form that is more useful to the consumer. With traditional information retrieval and filtering, documents themselves are typically unchanged by the processing.

The goal of information customization is to tailor an information artifact to a consumer's specific needs at a given moment. Since needs may change over even brief periods of time, the process tends to be inherently interactive. Customization benefits from facilities for browsing, extracting, abstracting, reporting, and so forth, especially when those facilities can be invoked interactively in real time as the consumer's needs change in response to the information being provided. Information filtering and retrieval, however, do not stress interactivity as a way to select artifacts.

We can also compare and contrast information customization with several other fields and techniques.

Hypertext and hypermedia. Hypermedia, typified by its subset hypertext, allows interactive navigation within a document or a set of documents. Hypertext browsing systems are not actually information customization systems because they do not transform documents into a customized form. Nevertheless, they are a step in that direction because they facilitate reading by allowing users to interactively determine the order in which information is presented.

While hypermedia and customization software both support nonlinear text presentation, in hypermedia the nonlinearity is *prescribed* by the author or authoring system. Thus, it is predetermined and static rather than dynamically tailored to the consumer. This prescriptive approach is a serious restriction. We challenge the hypothesis that restricting the freedom of information consumers is in their best interests. We also argue that the problem is not merely linearity of information but the prescriptive approach as well. Prescriptive nonlinearity imposes a predefined structure on information rather than allowing it to be structured according to user needs. Such a prescribed structure may not match the information consumer's current interests and objectives. An information customization approach avoids this unwanted imposition of predefined structure.

Hypermedia browsing is susceptible to the well-known "lost in hyperspace" phenomenon — the loss of one's contextual bearings and desired perspective. This can seriously compromise the utility of available information. While hypertext researchers are trying to alleviate this problem, nonhypertext approaches may be an important part of the solution. For example, a suitably generated customized extract can summarize a document from some point of view. A summary, by its nature, provides an overall perspective — albeit a perspective that may be idiosyncratic, depending on the customization requirements of the consumer.

Information extraction and knowledge discovery in databases. Information extraction from texts is typified by the third Message Understanding Conference (MUC-3). Participants fed news stories concerning Latin American terrorist incidents to their programs, which competed to effectively fill in a predefined framelike set of slots.[4,5] Successfully filled-in frames described key aspects of the news stories. Although useful, this does not constitute information customization because extraction occurred with respect to a predefined frame template provided months in advance.

The evolution of information technology calls for partial or even total automation.

Knowledge discovery in databases is related to information extraction and is typified by the Workshop on Knowledge Discovery in Databases,[6] where Ai et al. reported extracting chemical reactions from articles in the *Journal of Organic Chemistry*.[7] This extraction task used highly domain dependent properties of documents. Fonts, for example, are important clues in extracting chemical reactions from surrounding text. Knowledge discovery can also operate on nontextual information, such as computer-aided-design databases.[8]

MUC-3 and the work described at the Workshop on Knowledge Discovery in Databases represent high domain specificity as well as, typically, low interactivity. A natural extension to such extraction approaches would be to provide a significant degree of interactivity and flexibility in output, which could result in information customization systems.

Data interchange. Data interchange involves transforming information from one representation to another. Often, the same initial and target formats occur many times, as in converting documents from one word processing format to another. Sometimes, though, data must be converted to a specialized format for a one-time analysis. Typically, a program will be written to do this custom conversion. To our knowledge, no one has yet developed a tool for interactively helping a user specify initial and/or target formats. Such an artifact would be a useful information customizing tool.

Better tools are coming. Access to the "common world brain" envisioned by Wells will remain insufficient until information can be presented in a form customized to each consumer's evolving needs. Part 2, in the next issue of *Computer*, will provide some examples of information customization and describe our prototypes of information customizing systems.

References

1. H. Berghel, "Cyberspace Navigation," *PC AI*, Vol. 8, No. 5, Sept./Oct. 1994, pp. 38-41.
2. *Comm. ACM*, special section on information filtering, Vol. 35, No. 12, Dec. 1992, pp. 26-81.
3. N.J. Belkin and W.B. Croft, "Information Filtering and Information Retrieval: Two Sides of the Same Coin?" in Ref. 2, pp. 26-38.
4. B. Sundheim, ed., *Proc. Third Message Understanding Conf.*, Morgan Kaufman, San Mateo, Calif., 1991.
5. N. Chinchor, L. Hirschman, and D.D. Lewis, "Evaluating Message Understanding Systems: An Analysis of the Third Message Understanding Conf. (MUC-3)," *Computational Linguistics*, Vol. 19, Sept. 1993, pp. 409-449.
6. G. Piatetsky-Shapiro and W.J. Frawley, eds., *Knowledge Discovery in Databases*, MIT Press, Cambridge, Mass., 1991.
7. C.-S. Ai, P.E. Blower Jr., and R.H. Ledwith, "Extracting Reaction Information from Chemical Databases," in Ref. 6, pp. 367-381.
8. A.J. Gonzalez et al., "Automated Knowledge Generation from a CAD Database," in Ref. 6, pp. 383-396.

Dan Berleant is an assistant professor in the Computer Systems Engineering Department at the University of Arkansas.

Hal Berghel is a professor in the Computer Science Department at the University of Arkansas.

Correspondence can be addressed to either author. Berleant is at the Department of Computer Systems Engineering, 313 Engineering Hall, University of Arkansas, Fayetteville, AR 72701. Berleant's e-mail address is djb@engr.uark.edu; Berghel's e-mail address is hlb@acm.org

SOFTWARE

Software metrics:

Capers Jones,
Software Productivity Research

The software industry is an embarrassment when it comes to measurement and metrics. Many software managers and practitioners, including tenured academics in software engineering and computer science, seem to know little or nothing about these topics. Many of the measurements found in the software literature are not used with enough precision to replicate the author's findings — a canon of scientific writing in other fields. Several of the most widely used software metrics have been proved unworkable, yet they continue to show up in books, encyclopedias, and refereed journals. So long as these invalid metrics are used carelessly, there can be no true "software engineering," only a kind of amateurish craft that uses rough approximations instead of precise measurement.

Several widely used software metrics do not work, yet they continue to show up in books, encyclopedias, and refereed journals.

Software metrics that don't work. Three significant and widely used software metrics are invalid under various conditions: lines of code or LOC metrics, software science or Halstead metrics, and the cost-per-defect metric. The first two metrics are not invalid under all conditions, but they are when used to compare productivity or quality data across different programming languages. The third metric requires a