Towards Better Support for Copyright Compliance and for Privacy

Joan Feigenbaum
http://www.cs.yale.edu/homes/jf

The W3C Recommendation entitled "Architecture of the World Wide Web" (http://www.w3.org/TR/webarch) contains the following inspiring statement: "One goal of the Web, since its inception, has been to build a global community in which any party can share information with any other party." Two issues that continue to impede or at least complicate global sharing of information are copyright and privacy. Network-architectural decisions can go only so far in resolving these issues, because copyright and privacy are, to a large extent, social and legal (rather than technical) matters. Nonetheless, some architectural principles could facilitate sharing while respecting the goals of privacy and copyright compliance.

Copyright

In the arts and humanities, Web-based distribution of information and meta-information has greatly enhanced the ability of unaffiliated (or unfortunately affiliated) scholars and small scholarly organizations to participate in research and publication. Specifically, it has narrowed the opportunity gap between scholars with easy access to world-class (physical) libraries and museums and those without such access by making available on the Web digital representations of important works or searchable metadata about them. Google's plans to scan and make available for searching all of the books in several major research libraries is poised to accelerate this trend.

Unfortunately, copyright-related obstacles to full utilization of Web-accessible artifacts remain. Small scholarly organizations and independent scholars still find it difficult to identify all of the rights holders of works they need to use, and they find it impossible to negotiate terms from scratch. Technical standards for identification of rights holders and standard, nominal-fee licenses for scholarly and educational use should be deployed. From the general Google user's point of view, it is tragic that many library books may be left out of the searchable archive because publishers object to Google's earning ad revenue from their intellectual property without compensating them. A standard and simple way for Google and other search companies to share ad revenue with rights holders might break this logjam.

Privacy

As mass-storage devices get cheaper and cheaper, and networked computers become essential in more and more daily activity, a rapidly increasing amount of sensitive information about people and organizations is created, captured, stored, and mined. By "sensitive information," I need information that can harm data subjects, data

owners, data users, or other stakeholders if it is misused. This proliferation, dispersion, and longevity of sensitive information has led to widespread, justifiable anxiety about "privacy."

Before turning to consideration of architectural principles, I would like to endorse a linguistic change: The term "sensitive information" will probably be more productive in our attempts to solve these problems than the term "private information." The word "privacy" connotes secrecy or confidentiality; techniques such as cryptography and access control, developed to handle private information, are very effective at hiding information altogether from parties that are not authorized to use it and at making it available to authorized users. These techniques were developed in environments in which the sets of legitimate users were well defined, fairly small, and known at the time information was generated. None of these characteristics is true of the world of Web-based information sharing. As natural as it is to describe consumer-transaction records, medical records, or banking records as "private information," they may have to be accessed and used by large, dynamic, and unpredictable sets of people and machines in order to fulfill their reasons for being. Technical efforts should be focused on ensuring proper use of these sensitive data, rather than on hiding the data altogether from everyone whose need to use them cannot be foreseen.

There have already been some technical efforts focused on ensuring appropriate use rather than on enforcing secrecy, most notably the W3C's P3P project. These policy-language efforts should be continued, but they have an inherent limitation: Long-lived, general policies cannot capture enough of the dynamic human and technological context in which decisions are made. The policy approach should be complemented by technologies that support updating and correction of both sensitive data records and the policies that govern their use. Architectural principles that should guide technological development include:

- Support for transactions and interactions that do not result in long-term data storage. For example, many e-commerce websites give users the option of "proceeding as a guest" or "logging in to proceed." Users who proceed as guests should be able to rest assured that all of their personally identifying information will be discarded after the goods or services ordered have been delivered satisfactorily.
- Support for data-subject-driven quality control of long-lived databases. Periodically (but not frequently enough to be annoying), organizations that store sensitive information about people should allow those people to update and correct the stored data and to make new, well informed decisions about the policies that govern their use.
- Support for information provenance, history, and forensics so that problems and inaccuracies can be corrected recursively or at their sources.