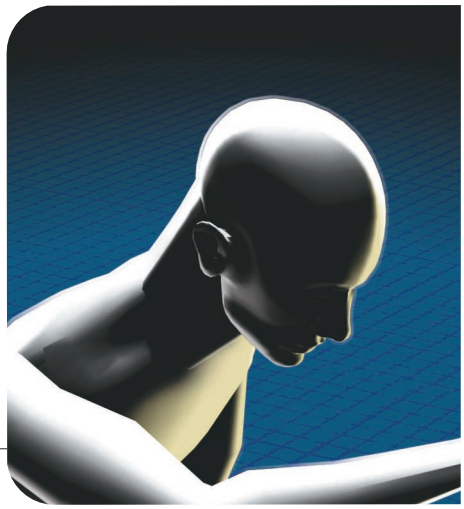
Financial Privacy Policies and the Need for Standardization

By analyzing 40 online privacy policy documents from nine financial institutions, the authors examine these important privacy notices' clarity and readability. Using goal-driven requirements engineering techniques and readability analysis, they found that compliance with the existing legislation and standards is, at best, questionable.



Annie I. Anton, Julia B. Earp, Qingfeng He, and William Stufflebeam
North Carolina State University

Davide Bolchini
University of Lugano

Carlos Jensen
Georgia Institute of Technology

The US Federal Trade Commission (FTC) states that a privacy policy should comprehensively describe a domain's information practices and be accessible on an institution's Web site by clicking on an icon or hyperlink.[1] The primary legislation covering financial data, the Gramm-Leach-Bliley Act (GLBA), also states that policies must be "clear and conspicuous." Yet, compliance with this GLBA requirement today is, at best, questionable. Although many organizations are taking strides to improve their privacy practices, and consumers are becoming more privacy-aware, it remains a tremendous burden for users to manage their privacy.

The purpose of privacy policies is to inform consumers about how organizations collect and use their personal information; they theoretically serve as a basis for consumer browsing and transaction decisions. Each policy differs greatly because of the lack of standardization across different industries and organizations. This lack of standardization for expressing business privacy practices presents a daunting learning curve for those wanting to compare different organizations' policies before deciding which organization to trust with their personal information. The lack of clarity in online privacy policy documents contributes to consumer privacy concerns, posing a serious impediment to expanded e-commerce growth and Internet usage.

We completed an in-depth analysis of 40 online privacy policy documents from nine financial institutions. We evaluated these privacy documents' level of clarity, intending to develop a standard vocabulary for expressing privacy statements, with the hope of increasing their clarity. Our analysis can also help policy makers and consumers identify practices that potentially threaten consumer privacy, and guide software engineers toward developing systems that are aligned with their organization's privacy policies.

The state of privacy policies

The Progress and Freedom Foundation (PFF) recently surveyed a random sample of highly visited Web sites and found that 83 percent posted a privacy policy,[2] showing a significant increase from the 1990s when only 14 percent provided any notice regarding information privacy practices.[1] Overall, the PFF study shows that commercial Web site privacy practices and policies have improved in two ways: they now claim to collect less information from consumers, and they increasingly reflect fair information practices.[3]

Although many organizations now post online privacy polices, they must realize that simply posting a privacy policy does not guarantee compliance with existing legislation. To date, US privacy protection law covers healthcare data (the Health Insurance Portability and Accountability Act, HIPAA), information obtained from and/or about children (the Children's Online Privacy Protection Act, Coppa), and financial data (the Gramm-Leach-Bliley Act, GLBA). The GLBA became effective on 1 July 2001. It requires financial institutions—including banks, insurance companies, and securities firms—to protect the security and confidentiality of nonpublic personal information (NPI) for distribution beyond the institution.

Although organizations only have to define one policy, users are expected to read the policy of every Web site with which they interact. Most users assume that others must have already read a Web site's privacy policy and identified any potential vulnerabilities. Moreover, these policies generally require that users have a reading equiv-

alency of at least two years of college education to fully comprehend them. Consequently, most Web site users do not attempt to read and understand the policies themselves.[4] Additionally, existing standards and tools meant to help end users manage their privacy, such as the Platform for Privacy Preferences Project (P3P, www.w3.org/P3P/) and Privacy Bird (www.privacybird.com), force users' preferences and concerns into defined categories, limiting their options.

Analyzing financial privacy policies

For our study, we examined 40 online privacy policy documents from nine GLBA-covered institutions (some institutions post multiple privacy-policy documents) using goal-driven requirements engineering techniques and text-readability metrics. Our sample consists of Web sites from three banks (Bank of America, Citibank, and Wachovia), three insurance companies (Allstate, American International Group, and State Farm) and three securities firms (Goldman Sachs, Merrill Lynch, and Morgan Stanley). We took this sample from a cumulative listing of the top five US banks (by revenue in 2002), the top five US property and casualty insurance companies (by net premiums written in 2001), and the top five US securities firms (by revenue in 2001). (These rankings are available at the Financial Services Fact Book's corresponding Web site, www.financialservicesfacts.org/.) The financial privacy policies we examined were in force during June 2003.

We used a content analysis technique called *goal mining*—extracting pre-requirements goals from post-requirements text artifacts[5]—to analyze Web site privacy documents. We conducted our goal-mining efforts in the spirit of Grounded Theory,[6] commonly used in qualitative research, in which existing phenomena are analyzed to understand the current state of a particular subject. Grounded Theory is derived from data that has been systematically gathered and analyzed. Therefore, the work we present here is not based on a distinct, preconceived theory or hypothesis that we hope to support or refute. Instead, our goal-mining effort was a scientific analysis to develop a new theory. The results of this kind of qualitative analysis should provide additional benefits to policy makers and consumers by providing more objective criteria for evaluating a Web site's privacy practices.

Previously, we successfully used the goal-mining heuristics that guide the process of extracting goal statements from policies to analyze nearly 50 privacy policies in general e-commerce and health-care Web sites.[5]

Mining policies

Goal mining refers to extracting goals from data sources (in this case, privacy policies)[7] by applying goal-based requirements analysis methods. The extracted goals are expressed in structured natural language. We begin the goal-mining process by first exploring any privacy policies to identify strategic and tactical goals. *Strategic goals* reflect high-level enterprise goals, whereas *tactical goals* involve short-term goal achievement. These goals are documented in a Web-based Privacy Goal Management Tool (PGMT) developed at North Carolina State University (NCSU).[8]

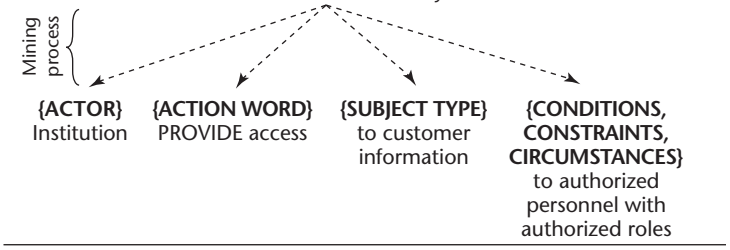


Figure 1. Privacy statement. Systematically decomposing a privacy statement lets us extract its semantics.

To identify goals, we analyzed each statement in a privacy policy by asking, "What goal(s) does this statement or fragment exemplify?" and "What goal(s) does this statement obstruct or thwart?" All action words are possible candidates for goals. Thus, we also identified goals in privacy policies by looking for useful keywords (verbs). The identified goals are worded to express a state that is true, or a condition that holds true, when the goal is realized.

Consider Privacy Policy 1 from the Bank of America privacy policy:

> "Employees are authorized to access customer information only when they need it, to provide you with accounts and services or to maintain your accounts."

By asking the goal identification questions, we identify the goal "G_{144}: PROVIDE access to Customer Information to authorized personnel with authorized roles" from Privacy Policy 1. (The careful reader will note that there is no explicit mention of an authorized role in Privacy Policy 1; we are interpreting the statement in the bank's favor for the purpose of this analysis, making the assumption about the definition of "need.") The goal's actor is the institution, and it is an integrity/security goal according to our taxonomy (which we describe later on).

This same goal was reused during the analysis of Privacy Policy 2 taken from the Citigroup Privacy Promise for Consumers:

> "We will permit only authorized employees, who are trained in the proper handling of customer information, to have access to that information."

Figure 1 shows how we decompose a privacy statement

Table 1. Summary of all Gramm-Leach-Bliley Act-covered financial institutions and respective privacy policies analyzed for this study.

	POLICY DOCUMENT	PROTECTION GOALS	VULNERABILITIES	UNCLASSIFIED GOALS	TOTAL	FLESCH READING EASE SCORE	FLESCH GRADE LEVEL
BANKS							
Bank of America	Overview	4	0	0	4	36.50	11.15
	Privacy policy	73	55	4	132	32.20	12.99
	Online practices	23	19	5	47	41.60	12.45
	Information security	29	2	12	43	39.10	11.68
	Identity theft	18	0	4	22	43.90	10.42
	Accounts and services	4	4	5	13	41.70	11.72
	FAQ: State NC	63	32	3	98	43.40	11.29
	Subtotal	214	112	33	359	40.10	11.77
Citibank	Citigroup Promise	21	15	0	36	24	15.65
	Citi.com Online Data policy	8	21	2	34	31.30	15.38
	Citi MyAccounts Promise	19	14	1	31	31.30	14.09
	Citi MyAccounts notice	12	14	1	27	30.10	15.54
	Citi terms of use	4	11	2	17	21.60	17.03
	Subtotal	64	75	6	145	28.50	15.47
Wachovia	Privacy Statement	60	28	10	98	29.70	13.32
	Internet privacy	20	40	4	64	35.20	13.91
	Privacy statement FAQ	44	25	7	76	30.20	13.25
	Fraud prevention	57	1	6	64	39.40	11.76
	Security statement	25	1	0	26	35.90	13.07
	Online banking and billpay	20	2	10	32	42.00	13.04
	Subtotal	226	97	37	360	35.00	12.95
INSURANCE COMPANIES							
Allstate	Privacy statement	55	23	11	89	41.10	11.68
	Terms of use	9	12	8	29	31.40	15.86
	Subtotal	64	35	19	118	38.10	12.89
American Int'l Group	Privacy policy	8	9	1	18	27.00	17.24
	Conditions of use	1	15	4	20	25.20	16.90
	Subtotal	9	24	5	38	26.10	16.97
State Farm	Privacy principles	3	2	1	6	23.10	14.93
	Privacy Policy customers	15	23	2	40	40.80	12.46
	Privacy Policy consumers	10	9	2	21	26.30	14.56
	Privacy and security	8	8	0	16	37.20	13.14
	Privacy policy for PHI	24	41	3	68	36.20	13.50
	State privacy rights	1	0	0	1	33.50	17.08
	Privacy policy FAQ	70	34	8	112	48.20	10.78

continued on next page

Table 1 continued.

	POLICY DOCUMENT	PROTECTION GOALS	VULNERABILITIES	UNCLASSIFIED GOALS	TOTAL	FLESCH READING EASE SCORE	FLESCH GRADE LEVEL
	Terms of use	11	8	11	30	31.00	13.98
	Subtotal	142	125	27	294	40.10	12.41
SECURITIES FIRMS							
Goldman Sachs	Privacy Policy	33	28	4	65	25.70	15.56
	Terms and Conditions of use	0	4	0	4	22.50	18.72
	Subtotal	33	32	4	69	24.10	17.15
Merrill Lynch	Global privacy pledge	30	34	5	69	30.90	14.84
	Online privacy statement	11	15	3	29	37.10	12.93
	Legal info	2	6	0	8	25.20	17.27
	Subtotal	43	55	8	106	29.50	15.55
Morgan Stanley	Privacy pledge	5	0	1	6	28.90	14.20
	US individual investor PP	23	41	3	67	28.90	15.76
	Internet security policy	8	10	0	18	31.40	14.79
	ClientServe ISP	18	3	3	24	35.90	13.30
	Terms of use	10	28	1	39	26.10	17.32
	Subtotal	64	82	8	154	29.50	15.70
Total		859	637	147	1643	33.07	14.1

into a privacy goal's four basic components during the goal-mining process. Each privacy statement decomposes into actor, action word, subject type, and conditions/constraints/circumstances. The *actor* represents the stakeholder responsible for achieving the goal; the *action word* represents the type of activity the statement describes; the *subject type* describes the kind of user information at issue; and finally, a goal usually recounts the conditions under which it actually takes place, the constraints to be respected, or other circumstances that provide the context to establish the goal's scope. We cannot entirely automate this process of discovery, decomposition, and representation because it requires significant semantic content analysis. This semantic analysis is best carried out by a team of analysts that does not simply decompose a goal into pieces but extracts each goal's meaning, exposing the relevant gist.

Tool support, however, can greatly enhance the mining process's efficiency. The PGMT, which assists analysts in the goal-mining, reconciliation, and management processes, supported our analysis. It maintains a goal repository for policy analyses and other documents from which we can derive goals. Each goal is associated with a unique ID, a description, a responsible actor, its sources, and a privacy taxonomy classification. All goals are fully traceable to the policies in which they appear and are distinguished as either policy goals (strategic goals) or scenario goals (tactical goals). We extracted 1,032 different goals from the 40 policies we examined using goal-mining identification heuristics. The PGMT tracked the number of goal occurrences in each policy (see column 6 in Table 1).

Clarity through reconciliation

Goal refinement in requirements engineering entails resolving ambiguities, redundancies, and inconsistencies that exist in the goal set, for eventual goal refinement into a requirements specification. Heuristics guide the goal-refinement process. For example, goals are considered synonymous if their intended end states are equivalent, or if they mean the same thing to different stakeholders who simply express the goal using different terminology. For example, the goals, "TRACK pages on our site using cookies" and "TRACK usage patterns using cookies" are synonymous, so we can reconcile them as one goal encompassing the spirit and scope of both. The analyst can choose either goal name; however, all related essential information must be maintained (such as actor and source). These two goals were ultimately merged with another goal, "TRACK usage patterns using aggregate data." We merged the previous two goals with the latter as follows: "G_{95}: TRACK usage patterns (using aggregate data or cookies)."

We also codified heuristics for how to reconcile hyperonymous (broader in meaning) and hyponymous

Common privacy policy keywords

ACCESS	CONNECT	DISCLOSE	MAINTAIN	POST	SHARE
AGGREGATE	CONSOLIDATE	DISPLAY	MAKE	PREVENT	SPECIFY
ALLOW	CONTACT	ENFORCE	MAXIMIZE	PROHIBIT	STORE
APPLY	CONTRACT	ENSURE	MINIMIZE	PROTECT	UPDATE
AVOID	CUSTOMIZE	EXCHANGE	MONITOR	PROVIDE	URGE
BLOCK	DENY	HELP	NOTIFY	RECOMMEND	USE
CHANGE	DESTROY	HONOR	OBLIGATE	REQUEST	VERIFY
CHOOSE	DISALLOW	IMPLY	OPT-IN	REQUIRE	
COLLECT	DISCIPLINE	INFORM	OPT-OUT	RESERVE	
COMPLY	DISCLAIM	LIMIT	INVESTIGATE	REVIEW	

(narrower in meaning) goals that are expressed using different keywords. For example, consider "G_{644}: PREVENT sharing customer information with telemarketers" and "G_{629}: AVOID selling/sharing customer information with telemarketers/other companies for marketing purposes." We applied two heuristics to these goals during goal refinement. These two goals were initially deemed synonymous with respect to their keywords. The keywords *avoid* and *prevent* mean to keep from happening. However, *prevent* implies an ensurance via a physical or procedural mechanism. In goals G_{629} and G_{644}, there was no mention of mechanisms to keep such sharing from happening, thus the keyword *avoid*. Additionally, because goal G_{629} is hyperonymous with respect to the argument of goal G_{644}, we maintained this goal (G_{629}). We made similar distinctions among different keywords. For example, the keywords *provide* and *inform* make something available. The distinction encoded in the PGMT is that *provide* refers to provision of services or specific functionalities, whereas *inform* refers to imparting knowledge or information.

Each keyword in the repository is formally defined to ensure consistent keyword use throughout the analysis process. The PGMT currently contains 57 keywords that are commonly found in Internet privacy policies (see the "Common privacy policy keywords" sidebar).

Additionally, we codified a rule set that suggests keywords used in different organizations' policies should be reconciled with those in the repository. As we just discussed, different institutions' policies express the same practices using different terms. These differences require end users to calibrate their understanding of different Web site policies, imposing a tremendous (and unfair) burden on them. It is not surprising, therefore, that end users find it difficult to trust the practices companies express in their policies.

From a systems perspective, a standard vocabulary would let us formalize the informal, ambiguous, and sometimes inaccurate statements found in Internet privacy policies. A standard vocabulary would also facilitate developing tools to help policy makers standardize their policies across institutions and help consumers more readily understand what is said in Web site policies. Tools employing such a vocabulary would make privacy policies more clear to consumers, letting them compare privacy practices concerning which companies they trust with their personal information. The 57 formally defined keywords in the sidebar provide a useful, extensible vocabulary for examining privacy policies because they standardize what different policies express with different terms in a manner that can increase end users' understanding.

After we completed this reconciliation process, the final goal set contained 910 goals extracted from financial policies.

Privacy protection or vulnerability?

Policies should express the ways in which nonpersonal information (NPI) is protected, but according to the Fair Information Practice Principles,[3] institutions should also inform their customers of potential vulnerabilities that could threaten privacy. Privacy protection goals express desired consumer privacy rights protections, whereas privacy vulnerabilities describe practices that potentially threaten consumer privacy. These two dimensions, protections and vulnerabilities, are extensively intertwined but not clearly separated in Web site privacy policies. However, it is important to help end users clearly distinguish between practices that protect their privacy and practices that introduce potential vulnerabilities. We developed a privacy goal taxonomy in which we broadly classify privacy statements as either privacy protection goals or privacy vulnerabilities.[5,7] The taxonomy provides a framework for understanding relevant privacy issues concerning how institutions treat customer data.

We subdivided privacy protection goals into five categories:[7]

- notice and awareness,
- choice and consent,

- access and participation,
- integrity and security, and
- enforcement and redress.

Notice and awareness goals reflect ways in which consumers should be notified or made aware of an organization's information practices before any information is actually collected from them. Choice and consent goals reflect ways in which organizations ensure that consumers are given options about what personal information is collected, how it might be used, and by whom. Access and participation goals reflect ways that organizations let customers access, correct, and challenge any data about themselves—for example, by providing a means for customers to ensure their data is accurate and complete. Integrity and security goals reflect ways in which organizations ensure data is accurate and secure. Finally, enforcement and redress goals reflect ways in which organizations enforce their policies.

Privacy vulnerabilities reflect existing threats to consumer privacy and represent statements of fact about behavior that might be characterized as privacy invasions, either obvious or insidious. Obvious privacy invasions are those that consumers are acutely aware of or about which they eventually become aware. Specifically, three kinds of obvious privacy invasions exist:

- direct collection for secondary purposes,
- personalization, and
- solicitation.

On the other hand, end users might not be aware of insidious privacy invasions, which include monitoring, storage, aggregation, and information transfer. Some might argue that if a consumer opts-in to being monitored, the following practices cannot possibly be insidious: having usage patterns or other data aggregated with that of other customers or having NPI stored in a database and/or shared with third parties. However, such information collection presents the potential for grievous privacy invasions simply because its existence creates vulnerability and, consequently, the potential for abuses.

We categorized all 910 goals according to these taxonomy classes by carefully considering each policy goal's intent. This let us compare the number of privacy protection goals and privacy vulnerabilities in each policy. Consider this goal: "STORE credit-card info securely (encrypted, separate database)." At first glance, the keyword *store* would have led us to classify this goal as a vulnerability (information storage). However, using the taxonomy let us identify this goal as a protection goal (integrity/security). Some goals were not truly relevant to privacy or privacy-related functionality and were unclassified for purposes of this study (see Table 1). For example, the Wachovia goal, "G_{548}: MAINTAIN efficient service," reflects the company's general declaration of commitment that expresses neither a privacy protection goal nor vulnerability.

For our analysis, we hypothesized that the number of protection goals in a financial institution's privacy policy would be greater than the number of vulnerabilities. We confirmed this by using a *t*-test analysis, which is a statistical technique used to compare two population means. When comparing the number of protection goals to the number of vulnerabilities in each financial policy (see Table 1), the *t*-test analysis revealed a statistically significant difference (p value = 0.02215) between them. In other words, we observed that the number of protection goals for a given Web site was, on average, greater than the number of vulnerabilities in that Web site. The p value, in particular, represents the chance that we would see such a large difference between the number of vulnerabilities and protection goals if there were only chance variability occurring. This was the case with 22 of the 40 examined financial Web site privacy policies. Although there were a higher number of protection goals in these policies, nearly 40 percent of the overall goals we found expressed a privacy vulnerability.

Identifying policy conflicts

Web-based systems' requirements, policies, and functionality are often misaligned or in conflict with one another.[7] This jeopardizes consumer privacy and makes it increasingly difficult for software developers to ensure their systems are secure and privacy-aware. One solution is to identify privacy policies' conflicts because they often reflect vulnerabilities that might otherwise go overlooked, resulting in unfortunate security breaches.

The goal-mining heuristics, coupled with the taxonomy, provide a basis for identifying conflicting statements in a privacy policy. To more vividly express the relationship between policy statements, we used the *i** notation, which supports modeling goals, their semantic relationships, and the corresponding stakeholders.[9] For example, the goal-mining activity identified a potential conflict between "G_{400}: MAINTAIN confidentiality of Customer Information" and "G_{401}: ALLOW offers from reputable companies" in the CitiGroup Privacy Promise policy (see Figure 2). Actually, G_{401} does not conflict with G_{400} per se, but it implies "G_{1165}: SHARE Customer Information with third parties," which strongly conflicts with G_{400}.

G_{400} is also in potential conflict with "G_{402}: OBLIGATE external companies to not retain Personally Identifiable Information [PII] unless customer expresses interest in their products/services." To elucidate the conflict, also in Figure 2, we split the goal into two parts: "G_{402a}: Retain Customer Information"—owned by the external company—and "G_{402b}: express interest in service"—owned by the customer. The policy suggests that G_{402a} is achieved only if G_{402b} is satisfied. However, the policy does not explain the criteria for determining customer interest in third-party services (see G_{402b}). This omission leads to ambiguity, which makes the conditions for which

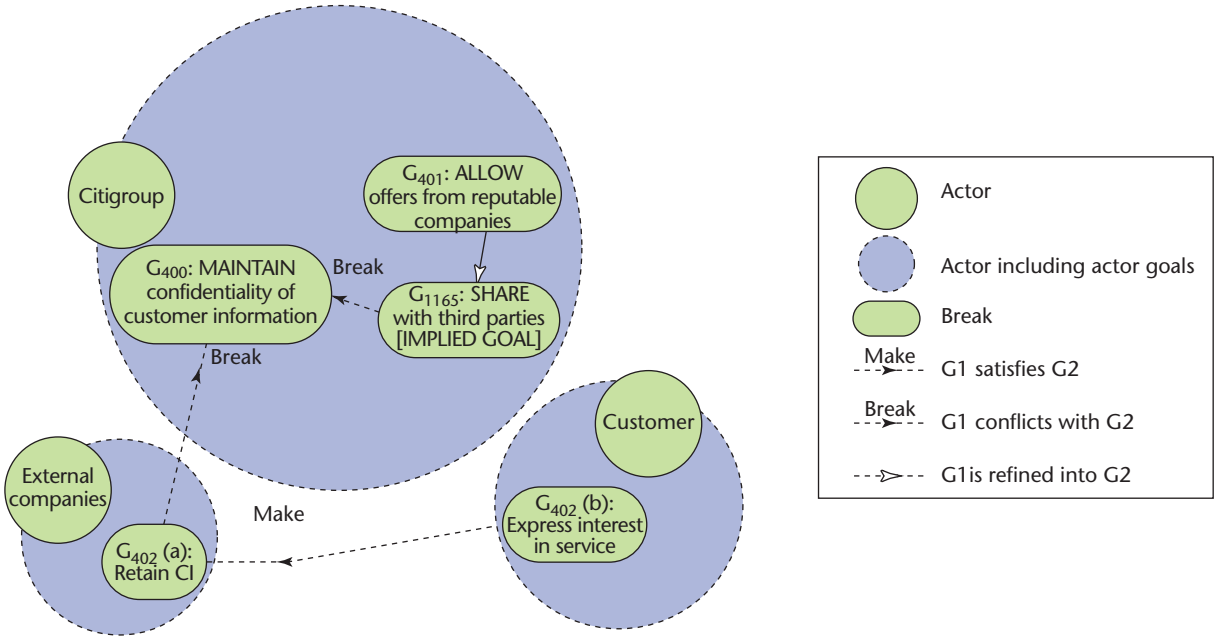


Figure 2. Potential conflicts in Citigroup Privacy Promise modeled with *i**.

customer information is retained (G_{402a}) by external companies unclear, thus compromising its confidentiality.

There are conflicts that are more evident than the one we show in Figure 2. For example, Bank of America's privacy policy had two consecutive goals: "G_{231}: HONOR customer privacy preferences once specified by customer" and "G_{232}: ALLOW six to eight weeks for privacy preferences to be fully effective." This situation presents an interesting question: How can Bank of America consistently honor customer preferences, when there is a six- to eight-week period during which they do not make those preferences effective? This is one of many questions that arise when considering such conflicts. Such questions are critical for software developers to consider because the implications are significant to end users who entrust their sensitive information to organizations' Web sites.

Customers should be aware that, when reading strong commitments in an institution's privacy policy, there might also be other easily overlooked potential vulnerabilities that undercut those commitments. Goal-based analysis exposes these kinds of conflicts that an end user might not discern with a superficial reading. This form of analysis lets policy makers gain deeper insights into the semantic relationships that form the meaning of their current policy. Based on these insights, policy makers can devise strategies to resolve conflicts and inconsistencies, resulting in higher quality policy statements that better conform to the GLBA's requirement for clarity.

Are privacy policies readable?

The GLBA's clear and conspicuous policy requirement means that institutions should post "a notice that is reasonably understandable and designed to call attention to the nature and significance of the information in the notice" (see 16 C.F.R. Part 313.4(a)). This clarity requirement is critical because consumers are interested in knowing how to protect their privacy. Hence, we examine privacy policies by looking at their clarity and readability.

What constitutes a clear notice hinges on the language used in it and whether it is reasonable to expect the target audience to understand it. Although the GLBA provides useful examples, they are not exhaustive, so it is largely up to financial institutions to make subjective judgments about their notices' clarity. A more objective measure considers the target audience's reading and comprehension skills, as well as the notices' readability, to determine whether they are clear enough to be understood.

Literacy and education are closely linked to income, and as computers and Internet access are still relatively expensive, we expect the population of US adults who are online to have a higher than average education and literacy rate. Thus, we use the education level statistics for the US adult Internet population rather than that of the general population (see Table 2).

We know from the 2000 US Census that 15.5 percent of the US population over the age of 25 has less than a high school education and only 26.9 percent of the population has a bachelor's degree or higher.[10]

Additionally, 65.6 percent of the US population has access to a computer, and 53.9 percent is now online.[10] The adult US population's average education is 13.5 years, whereas the Internet population has an average

Table 2. Educational level and Internet use.

EDUCATIONAL LEVEL	GENERAL POPULATION (GP)		PERCENT GP ONLINE	INTERNET POPULATION	
	NUMBER OF PEOPLE (IN MILLIONS)	PERCENT OF TOTAL POPULATION		NUMBER OF PEOPLE (IN MILLIONS)	PERCENT OF INTERNET POPULATION
Less than high school	27.5	15.5	12.8	3.5	3.8
High school diploma/GED	57.4	32.4	39.8	22.8	24.5
Some college (Associate's degree)	45.4	25.6	62.4	28.3	30.5
Bachelor's degree	30.6	17.7	80.8	24.7	26.6
Beyond bachelor's	16.3	9.2	83.7	13.6	14.6

Source: 2002 National Telecommunications and Information Administration report[10]

14.4 years. (Average calculated with the following values: All with less than high school education equal 11th grade, high school equals 12th grade, some college equals 14th grade, college equals 16th grade, and postgraduate equals 17th grade.) Although Internet users, on average, are more educated, 28.3 percent of adult US Internet users have the equivalent of a high school education or less.

With a greater understanding of the population's literacy level, we can examine whether privacy notices are clear, as the GLBA requires. The most commonly used method employs a standardized, statistical readability metric that allows objective evaluation and simple comparison between notices. The Flesch Reading Ease Score (FRES)[11] is a metric for evaluating complex texts that is often used to evaluate school texts and legal documents. FRES gives an approximate score for a text's difficulty. Although no metric is universally accepted, the FRES metrics have been commonly accepted benchmarks for decades. For example, the FRES is used by state and federal regulators to judge the complexity of insurance policies in more than 16 US states.

The FRES rates texts on a 100-point scale, on which higher scores signify simpler texts. We calculate this score by looking at the average number of syllables per word and the average sentence length. Longer words and sentences are more difficult to read and therefore produce a lower FRES. The Flesch Grade Level (FGL) determines the US grade-school equivalency level of a text and is also based on the average number of syllables and sentence length. We compute the metrics as follows:

- FRES: 206.835 – 84.6 × (total syllables/ total words) – 1.015 × (total words/total sentences)
- FGL: (0.39 × average sentence length (in words)) + (11.8 × average number of syllables per word) – 15.59.

Several tools calculate the FRES, including Microsoft Word, which we used to evaluate the policies we discuss here. MS Word also calculates the FGL up to the 12th grade; we calculated the scores for more complicated texts manually using the Flesch formulas. Table 1 lists the scores for the policies we examined.

Our survey found the average FRES to be 33.1 (standard deviation of 6.8); the average grade level required to read these policies is 14.1 (standard deviation of 2.1). This average is lower than the average education level of US adults who are online, but it is higher than that of the general population. The most difficult policy had an FGL of 18.7 (Goldman Sachs terms and conditions of use), the equivalent of a postgraduate education. The most readable notice (as rated by the FGL) required the equivalent of 10.4 years of schooling to understand (Bank of America's identity theft).

Of the 40 policies examined, eight require the equivalent of a high school education or less (12 years), 13 require the equivalent of some college education (12 to 14 years), 12 require 14 to 16 years of schooling, and seven require the equivalent of a postgraduate education (more than 16 years). Moreover, of the nine institutions in our sample, six had at least one policy document requiring the equivalent of a postgraduate education. This means that even though the average grade level equivalent is 14.1, a full understanding of what two-thirds of these organizations are doing is perhaps only available to one-sixth of those US adults who are online.

The fact that 28.3 percent of adults are likely unable to understand these privacy policies challenges the principle of clear notice; none of the nine institutions examined live up to the GLBA clarity requirement's intent. Bank of America comes close to that with two of its seven notices requiring more than a high school education to understand, but none require more than 13 years of schooling. Setting the bar more leniently at the equivalent of a college education (excluding 58.8 percent of the population), only three of the nine organizations examined (Allstate,

Wachovia, and Bank of America) give appropriate notice. In either case, we are clearly a long way from meeting this GLBA requirement.

Applying our goal-mining heuristics in conjunction with our privacy protection and vulnerability taxonomy yields results that organizations and institutions can effectively use in several ways. First, both vendors and customers can use the mining process we describe to systematically analyze and compare competitors' Internet privacy policies, and thus better understand the semantic structure of their current policy documents. Second, we can explore the scope of the privacy practices that the standardized keyword set identifies and compare them to the statements in an organization's privacy policy. Policy makers can use this to assess their policies' coverage with respect to the various security and privacy considerations that must be taken into account. For example, some organizations might not address important issues such as managing customer preferences in the privacy policy—something that the analyzed institutions' policies often neglected—or recommendations to the users about his or her online behavior. The keyword set and the examples we provide in this article suggest techniques to ensure coverage of critical security and privacy requirements.

An important result of our study is the privacy taxonomy, which policy makers can use to classify their privacy statements as either privacy protection goals or vulnerabilities. This classification can help policy makers develop a clear picture of their privacy practices' communication orientation—for example, do they mainly protect or threaten consumer privacy? Finally, the *i** notation we use can help policy analysts easily represent semantic relationships among privacy statements by abstracting the relevant elements from the maze of available documentation.

Even though we examined policies from a single domain, it was challenging to understand what the different policies meant because certain statements used distinct vocabularies despite the GLBA's existence, which advocates standardization. This vocabulary difference required us to spend a great deal of time recalibrating our understanding during the goal-refinement activities. What is perhaps most concerning is that if experienced privacy policy analysts encountered difficulties in understanding various policy statements (even with tool and methodological support), we can expect customers to continue to face difficulties if change does not occur. Our research shows that institutions must standardize how they express their privacy practices and that online privacy policies can be more clear, benefiting both the institutions and end users.

Additionally, our preliminary analysis has revealed ambiguities and conflicts that we are now analyzing to codify rules for their identification and subsequent resolution. For example, some policies contain temporal statements that are simple to check for conflicts because they establish constraints within which other policy statements must comply. Our analysis thus far has been limited to identifying conflicts and ambiguities in each privacy document. However, a need exists for additional, systematic ways to identify conflicts and ambiguities not only in a single policy but also across an organization's various privacy documents. Having many policies and goals per institution requires additional analysis, and such an analysis is the focus of our research's next phase.

Privacy will continue to be important to consumers. We believe clearly articulated, meaningful privacy policies are also important to good business practice. We continue to investigate methods to support the goals of all stakeholders with an interest in clear privacy practices. □

Acknowledgments

The National Science Foundation Information Technology Research (ITR) grant #0113792 and a doctoral grant from the Swiss National Science Foundation supported this work. We thank Colin Potts; Thomas Alspaugh for his participation in the goal-mining process; and Gene Spafford for his comments on this article.

References

1. Federal Trade Commission, *Privacy Online: A Report to Congress*, Jun. 1998, www.ftc.gov/reports/privacy3/.
2. W.F. Adkinson, J.A. Eisenach, and T.M. Lenard, *Privacy Online: A Report on the Information Practices and Policies of Commercial Web Sites*, Progress & Freedom Foundation, 2002; www.pff.org/publications/privacyonlinefinalael.pdf.
3. US Dept. of Health, Education, and Welfare, *The Code of Fair Information Practices*, Secretary's Advisory Committee on Automated Personal Data Systems, Records, Computers, and the Rights of Citizens, 1973, vol. viii; www.epic.org/privacy/consumer/code_fair_info.html.
4. J.B. Earp and D. Baumer, "Innovative Web Use to Learn about Consumer Behavior and Online Privacy," *Comm. ACM*, vol. 46, no. 4, 2003, pp. 81–83.
5. A.I. Antón and J.B. Earp, "A Requirements Taxonomy to Reduce Website Privacy Vulnerabilities," to be published in *Requirements Engineering J.*, 2004.
6. B.C. Glaser and A.L. Strauss, *The Discovery of Grounded Theory*, Aldine Publishing Co., 1967.
7. A.I. Antón, J.B. Earp, and A. Reese, "Analyzing Web Site Privacy Requirements Using a Privacy Goal Taxonomy," *Proc. 10th IEEE Joint Requirements Eng. Conf.*, IEEE CS Press, 2002, pp. 605–612.
8. N. Jain et al., *The Privacy Goal Management Tool (PGMT) Software Requirements Specification*, tech. report #TR-2004-7, North Carolina State Univ., Computer Science Dept., 2004.
9. E. Yu, "Modeling Organizations for Information Systems Requirements Engineering," *Proc. IEEE 1st Int'l Symp. Requirements Eng.*, IEEE CS Press, 1993, pp. 34–41.
10. Nat'l Telecommunications and Information Administra-

tion, *A Nation Online: How Americans Are Expanding Their Use of the Internet*, 2002; www.ntia.doc.gov/ntiahome/dn/.
11. R. Flesch, *The Art of Readable Writing*, Macmillan Publishing, 1949.

Annie I. Antón *is an associate professor in the North Carolina State University College of Engineering, where she is a Cyber Defense Lab member and director of The Privacy Place (the privacyplace.org). Her research interests include software requirements engineering, information privacy and security policy, software evolution, and process improvement. She has a BS, MS, and PhD in computer science from the Georgia Institute of Technology. She is a member of the ACM, the IAPP, and a senior member of the IEEE. Contact her at 900 Main Campus Dr., Ste. 165C, Raleigh, NC 27695-8207; aianton@eos.ncsu.edu.*

Julia Earp *is an assistant professor of information technology in business management in the North Carolina State University College of Management, where she is codirector of the E-Commerce Studio and a member of The Privacy Place. Her research focuses on Internet security and privacy issues from several different perspectives, including data management, consumer values, systems development, and policy. She has a PhD in information technology from Virginia Tech. She is a member of the ACM and the IEEE. Contact her at North Carolina State Univ., Raleigh, NC 27695-7229; Julia_Earp@ncsu.edu.*

Davide Bolchini *is a research and teaching assistant in new media design at the Faculty of Communication Sciences of the University of Lugano and a lecturer of usability at the Politecnico di Milano (Como campus). His research focuses on Web requirements analysis, usability evaluation, and conceptual design. He has a PhD in communication sciences from the University of Lugano. He is a member of the Usability Professional Association, the ACM, the IEEE, and Association for the Advancement of Computing in Education. Contact him at the TEC-Lab, Faculty of Communication Sciences, Univ.of Lugano, via G.Buffi 13 – TI 6900 Lugano, Switzerland; davide.bolchini@lu.unisi.ch.*

Qingfeng He *is a PhD candidate in the Computer Science Department at North Carolina State University, where he is a member of the NCSU Cyber Defense Lab and The Privacy Place. His research interests include software requirements engineering, security and privacy requirements, and policy specification. He has a BE and ME in electrical engineering from Tsinghua University, China. He is a member of the ACM. Contact him at 900 Main Campus Dr., 165C-197, Raleigh, NC 27695; qhe2@eos.ncsu.edu.*

Carlos Jensen *is a PhD candidate in the College of Computing at the Georgia Institute of Technology, specializing in human–computer interaction and privacy. His research focuses on ways to make privacy and security visible and manageable to users. He has a BS in computer science from the State University of New York, at Brockport. He is a member of the ACM and the IEEE. Contact him at the College of Computing, Georgia Inst. of Tech., Atlanta, GA 30332; carlosj@cc.gatech.edu.*

William Stufflebeam *is a PhD student in the Computer Science Department at North Carolina State University, where he is a member of the Cyber Defense Lab and The Privacy Place. His research interests include requirements coverage, information security and privacy, and applying requirements engineering practices in multidisciplinary efforts. He has a BS in management information systems from North Carolina State University and is a member of the ACM. Contact him at 900 Main Campus Dr., 165C-197, Raleigh, NC 27695-8207; whstuffl@eos.ncsu.edu.*