Best practices and tools for personal information compliance management

M. Kudo
Y. Araki
H. Nomiyama
S. Saito
Y. Sohda

Recent incidents involving the loss of personal information and identity theft have raised concerns worldwide over information privacy. In Japan, the Personal Information Protection Act went into effect in April 2005, requiring every enterprise to manage sensitive personal information on servers, workstations, and personal computers throughout the organization. This paper describes two tools we developed to assist in the management of personal information, aDesigner and the Personal Information Detection (PID) tool. The aDesigner tool scans an entire Web site to determine if each HTML page complies with the IBM privacy guidelines for external Web sites. PID is capable of automatically identifying "named entities," such as personal names, addresses, or telephone numbers in the textual parts of target files based on Japanese morphological analysis technology. This paper also summarizes the best practices used in IBM Japan for privacy management and presents statistical results concerning personal information gathered through deployment of these tools.

INTRODUCTION

Privacy concerns have been rapidly increasing because of repeated incidents involving the loss of personal information, user identity theft, and unexpected personal information leakage. For example, the ***Boston Globe*** reported in 2005 that the Bank of America lost tapes containing personal financial information for 1,200,000 accounts belonging to federal employees.[1] The Federal Trade Commission released a survey of identity theft in the United States in 2003, stating that there were 27.3 million victims in the prior five years.[2] As the frequency of these incidents shows, it is no longer an easy matter for an enterprise to protect privacy in conventional ways in the computer age. Governments and companies are responsible for implementing effective countermeasures to reduce the number of such incidents.

In Japan, the Personal Information Protection Act[3] (PIPA) took full effect in April 2005, heightening concern about ways to protect the personal information of customers and employees. In PIPA, personal information (PI) is defined as information

that identifies a specific living person. Anonymous information and corporate information are not included because they do not enable identification. Basic PI is the person's name, address, date of birth, and gender, but PI includes other items such as telephone number, e-mail address, driver's license number, employee number, purchase history, and so forth. PIPA imposes obligations on the handling of PI by business operators. If a business operator holds the PI of more than 5000 people at any time for six months after obtaining that information and uses it to conduct business, legal obligations to properly handle the PI are imposed on the operator. Business operators in violation of this act receive administrative guidance from the responsible minister of the supervisory agency (the Ministry of Economy, Trade, and Industry, for example). Such guidance may include compulsory reporting, advice, business improvement recommendations, or administrative orders. If the operator fails to comply with such recommendations or orders, the operator can be punished by up to six months in prison or by a fine of up to 300,000 Japanese yen (equivalent to about $2,800). Since the act went into effect, there have been many complaints from affected individuals, some compensation paid to victims, a few claims for damages under the Civil Code, and many media reports about privacy and security incidents.

Since 2005, Japanese companies have been preparing to comply with PIPA. According to the "Survey on Communication Usage Trends"[4] reported by the Japanese Ministry of Internal Affairs and Communications, 73.2 percent of companies had implemented some compliance measures by March 2006, a significant increase from only 56.5 percent the year before. In particular, for companies with 500 or more employees, almost 90 percent had taken action. In addition, almost half of the companies had already implemented employee education programs or created an organizational role that is responsible for PI management in the company. Despite these efforts by many Japanese companies seeking to prevent PI leakages, there are still many incidents happening even in large companies, often caused by computer viruses in an employee's PC that maliciously disclose PI stored in the PC over the Internet.

The critical problems in privacy management are twofold: (1) PCs owned or used by employees are likely to contain sensitive and confidential information, which is often out of the scope of the internal governance process; and (2) companies need to be aware of the current detailed status of how PI is handled, stored, and protected in the company.

Some companies have introduced PCs without hard disk drives, known as *thin clients*, to help their employees overcome the first of these problems. Since every user file is stored and managed in a centralized way, it is much less likely for employees using thin clients to mishandle PI. However, this type of solution requires sufficient financial support to replace the current IT assets and makes dramatic changes in the IT environment of the company, which is sometimes unrealistic. Thus, approaches are needed that work in current IT environments. The second problem requires a security inventory of PCs, workstations, host systems, and their applications throughout the company, which requires active support from application owners, administrators, and employees. We are not aware of any statistical data on such security inventory results.

The goal of this paper is to provide a holistic lifecycle model for PI and describe related tools to assist in the internal governance process without introducing new IT assets such as thin clients. This paper also presents best practices being used in IBM Japan and statistical results concerning how PI is stored and managed.

Privacy policies in IBM Japan

IBM has more than 30 years of experience in privacy protection and has been a pioneer in terms of initiatives on privacy policy development and in instituting the position of a chief privacy officer (CPO). In IBM, a global privacy policy (for the protection of personal information) is applied to every process and application. This protection applies to both employee and customer information. The key elements of this approach are: (1) fair and lawful collection and use of information, (2) disclosure of the reasons for collecting the data at the time of collection and use of the information only for the purposes disclosed, (3) assurance of the accuracy of the data, (4) restrictions on internal and external use of information, (5) technical and organizational security measures and proper supervision of subcontractors, and (6) appropriate responses to requests for information disclosure.

IBM's privacy policy specifies general principles underlying internal rules for the collection, use,

retention, disclosure, and supply of PI in IBM. These policies are consistent with PIPA. IBM also provides global internal rules for privacy and security concerning PI. For security, the rules specify the IT security of PCs and servers, protection of confidential information, control procedures for entering and leaving facilities, and anti-disaster measures. For privacy, the rules provide guidelines for the protection of customer PI, the design of Web pages for online collection of PI, the distribution of marketing e-mail, and the protection of employee PI. Implementation of the IBM guidelines is the responsibility of each division. Observance of the IBM guidelines is the duty of each employee.

Most of the IBM Japan group companies fall under the PIPA definition of business operators handling PI. This includes not only customer information on individual people, but also information on the corporate employees, employees of vendors for purchases or business partners, employment applicants, and employees of the IBM Japan group. Anonymous information and corporate information is not included because it does not make identification possible.

A leakage of PI can cause very serious risks for the company. When a leakage of PI occurs or is suspected, the relevant department must immediately ask for support from the CPO's office for adequate countermeasures. Responding to such a report, the CPO will assess the situation and give advice, but if the extent of the problem exceeds the scope of the CPO's authority or cannot be dealt with by ordinary processes, or if the problem may cause a managerial crisis, the matter comes under the corporate-wide crisis management system.

Privacy policy deployment

Defining a company privacy policy according to the pertinent laws and regulations is very important, but it does not necessarily mean that the specified privacy policy is implementable on the company's existing IT systems. In some cases, human-readable privacy policy can be interpreted in several different ways. It needs to be refined and instantiated for a specific region and for specific IT systems. We illustrate this with two privacy policy instances.

Policy for PI collection using Web interfaces

In general, customers can visit IBM on the Internet without identifying themselves or giving PI. There are times when IBM may need information from them; for instance, to process an order, to correspond, to provide a subscription, or in connection with a job application. In these cases, all of the PI is handled in accordance with "IBM Privacy Practices on the Web,"[5] a Web page to which all of the externally accessible IBM Web pages are linked. It states that IBM requires that customers be informed as to how IBM will use such information before it is collected from them; if customers specify that they do not want IBM to use this information, for example, to make further contact with them beyond fulfilling their requests, IBM respects their wishes.

IBM internally defines the privacy policy for PI collection applied to every externally accessible Web site. The policy consists of seven items divided into two categories, one concerning notifications to customers and the other concerning security issues.

Customers must be informed of the usage of the collected PI so that they have the option to cancel its input, to restrict its usage, and to specify the ways in which IBM may contact them. The policy also requires that all external IBM Web pages must include a link to the general IBM Privacy Practices statement. From the security viewpoint, it states that communication between the customer and the Web server should be appropriately encrypted and secured in the way specified in the policy.

Definition of highly sensitive PI

In IBM Japan, the privacy office has defined a category of highly sensitive PI, which includes PI of clients, the corporate staff, and employee PI that contains sensitive information such as human-resource and health-care information. Because such PI is very sensitive, the privacy office asks every employee to make sure that any files containing such PI are handled in an appropriate manner. A typical way to handle them is to delete them if not necessary, to encrypt them using passwords, or to move (not copy) those files into another secure computing environment (e.g., a database).

Related work

In this section, we describe work related to the topics in this paper. In particular, we summarize related work in the areas of privacy standards and tools for privacy management, PI detection, and privacy compliance checking.

Privacy standards

Standardization of privacy procedures has taken a diverse number of forms. The P3P (Platform for Privacy Preferences) specification[6] is a W3C** (Worldwide Web Consortium**) standard for communicating privacy policies (e.g., purpose, opt-in and opt-out policies, etc.) between a hosting server and a consumer using a Web site. It defines the vocabulary and syntax of the privacy policy, but policy enforcement (either by the client PC or by the hosting server) is outside its scope.

The XACML (Extensible Access Control Markup Language) policy language[7] can be used as a privacy policy specification language. The specification defines a "policy enforcement point," although it is not positioned for PI life-cycle management nor for existing IT systems. It is possible to specify privacy policies in XACML that are subsumed by the three policy enforcement points described in this paper.

Privacy management tools

Backes et al. propose a tool for management of enterprise privacy policies.[8] Their proposal includes the syntax and semantics of the privacy policies, refinement and auditing of the policies, and the composition of two or more policies. Enforcement of the privacy policies at the system level is outside the scope of their paper.

Ashley et al. propose an architecture for enterprise privacy management and enforcement.[9] Their approach is similar to ours in the sense that their architecture has multiple enforcement points where each data access is monitored according to the corresponding privacy policies. The difference is that their primary target for the IT environment is a server platform, whereas ours targets both server and client environments. In addition, their proposed tool requires preliminary mapping between a policy vocabulary and application semantics, whereas our tool does not need such preliminary work before deployment and use.

Giblin et al. propose REALM (Regulations Expressed as Logical Models), a metamodel for modeling regulations and managing them in a systematic life cycle.[10] Their focus for the life cycle is a policy model where operational semantics are formally expressed. In contrast, the current paper focuses on the life cycle of the PI data. Thus their policy life-cycle model complements our proposed PI life-cycle model.

Numao et al. have published a technology preview that provides a Java** library for adding privacy policy enforcement to existing Web applications which use JDBC** (Java Database Connectivity) / SQL (Structured Query Language).[11] It provides privacy monitoring functions for applications running on IBM WebSphere* Application Server and connecting to IBM DB2* by means of JDBC. It also provides a specific implementation for the target-software enforcement point (TS-EP) that is used in our model. The current paper presents a PI life-cycle model which positions such enforcement points using a holistic view.

Tivoli* Security Compliance Manager[12] provides a policy compliance architecture consisting of a server-side compliance management point and a client-side monitoring point. Certain parts of our PI management architecture could be implemented using the TSCM functions.

PI detection tools

Commercial products for detecting PI in Japanese texts are available.[13–16] Some of these tools detect PI by regular-expression pattern matching and keyword matching of person and place names, without linguistic analysis.[17] Linguistic analysis is effective for precise detection (to avoid over-detection and missing sensitive documents) because some words in Japanese are ambiguous in their word types. For example, the word "出口 (Deguchi)" is a personal name (specifically, a family name) and also a general noun (meaning "exit"). If "出口" is registered as a personal name in keyword matching, many documents which include the general noun "出口", would be detected. By using linguistic analysis, only documents in which the personal name "出口" exists will be detected.

Hosomi et al. propose a mechanism using "slot filling" to detect sensitive documents through textual and structural analysis.[18] The method calculates the concentration of elements of sensitive information, such as persons' names, phone numbers, and so on, to identify sensitive information more precisely. It does not consider whether sensitive information is sufficient to identify people, which may result in over-detection of PI. Slot filling is a simple and classical approach in artificial intelligence research and is effective for recognizing PI.

Privacy compliance-checking tools

Watchfire WebXM[19] is an online risk management solution to audit Web sites for issues impacting compliance and effectiveness. It has different add-on modules to process each kind of issue, one of which is the privacy module. It identifies the data collection forms on the Web site, specifically those collecting PI. It reports on the security settings of these forms and checks to ensure that each has a link to the appropriate privacy policy. It also checks for compliance with the relevant laws and regulations. While the WebXM privacy module is a general compliance checking tool for privacy standards, our proposed checker has more flexible and accurate checking functionalities that were needed to enforce IBM's privacy policy.

In the remainder of this paper, we propose a PI life-cycle model that combines the PI state and the enforcement points in the IT environment. A typical instantiation of the PI life-cycle model is presented. We then describe the aDesigner and PID tools, both of which effectively assist companies in complying with PIPA. Statistical results derived from the application of these tools in IBM Japan are presented.

PI LIFE-CYCLE MANAGEMENT

In this section, we present a data-centric view of PI. We then propose a PI life-cycle management model (PI-LM model) which enforces practical and effective PI protection processes by using the data-centric view of PI.

State of digital PI data

There are four typical states of PI data. First, PI data is *created*, for example by input from the user or by Internet download of a file. Next, PI data is *consumed* by other entities for some purpose, such as order processing or Web site improvements. PI data is sometimes *updated* by the person who retains the data. The consumption and the update processes may occur arbitrarily many times in the life cycle. Lastly, PI data is *terminated* when the file system, application, or database deletes it. We call these four states and the transitions between them a PI-data life cycle.

As PIPA states, each enterprise should provide an appropriate and effective scheme to protect PI according to its privacy policy. Because the privacy policy often includes high-level goals in a human-understandable format, there is a wide gap between an effective protection scheme at the system layer and the privacy protection policy. Many research efforts have attempted to fill this gap.[10,20] The challenge in doing so can be summarized as follows. A privacy policy should be specified in an IT-level language or by use of low-level expressions, configurations, or executable programs at the system implementation layer; the protection system should provide several policy enforcement points that enforce the privacy policy at the implementation layer; and the protection system should also provide a compliance-checking monitor to determine if the privacy policy has been satisfied.

There are many known approaches to address these issues. These approaches are applied at separate enforcement points. For example, it is straightforward to map the opt-in privacy policy to an HTML (hypertext markup language) checkbox which allows users to choose how their PI may be used. An operating-system-level logging facility can be used to address the issue of monitoring without adding a new enforcement point to the application. In this paper, we focus on the enforcement and monitoring issues in terms of the life cycle of the PI data.

PI life-cycle management model

We propose the PL-LM model, which combines each state of the PI data with four policy enforcement points. ***Figure 1*** illustrates the proposed PI-LM model, consisting of the PI management component, PI policy enforcer (PI-PE), and the target software domain.

Management component

The PI management component consists of the PI-policy management unit and the PI-state management unit. Both are primarily used and managed by the CPO. The PI policy enforcer (PI-PE) performs compliance checks on a particular target software domain according to the PI policy specified by the PI management component. The PI-PE handles the life cycle of each unit of PI data whenever PI data (1) enters into the target software domain, (2) is consumed and updated, or (3) is terminated (deleted). From a practical viewpoint, the PI-PE is instantiated at each enforcement point, such as a file filter driver, a database query rewriter, or a file scanner. The results of the compliance check are returned to the PI-state management unit, where

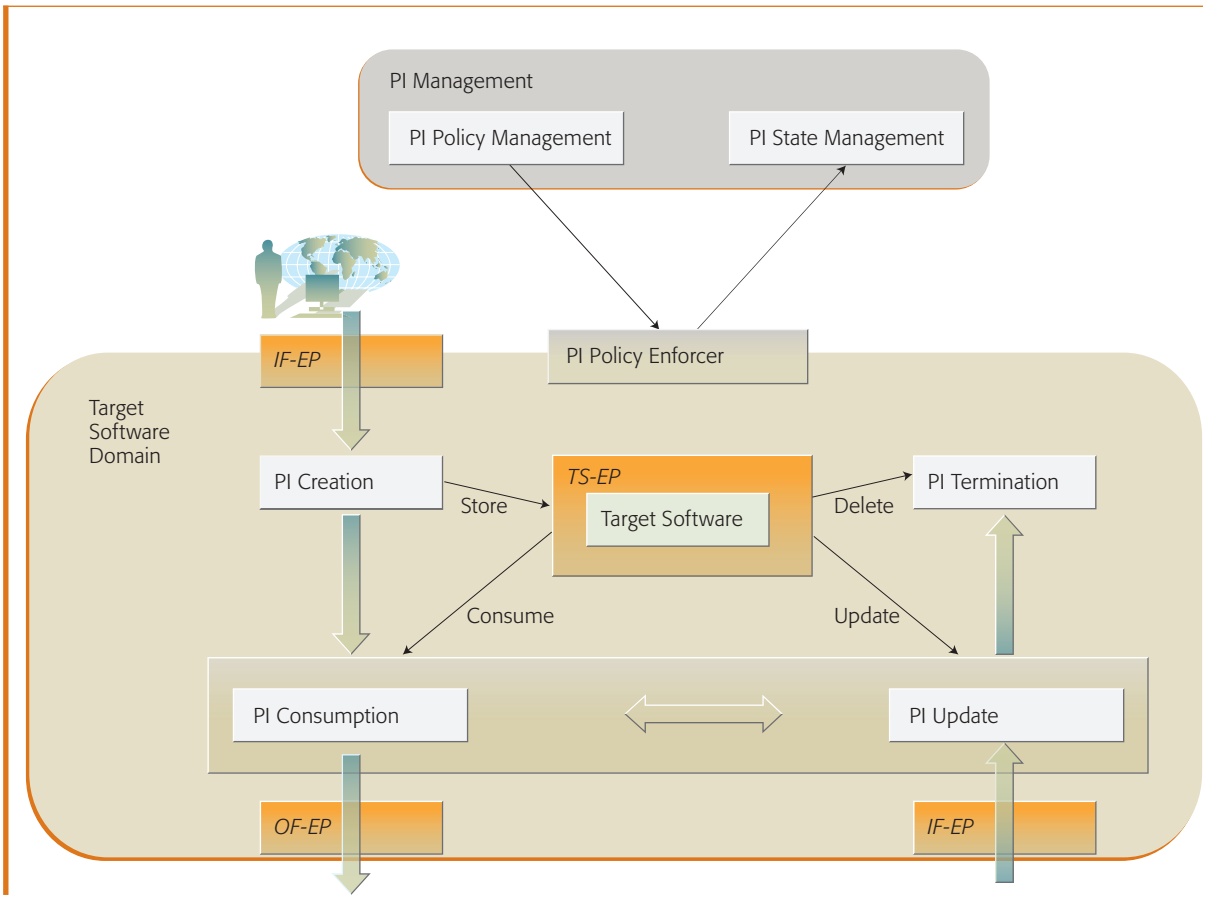

Figure 1
PI life-cycle management model

the CPO analyzes the current state and calculates the associated risk.

Policy enforcement points

As shown in Figure 1, there are three kinds of enforcement points at four different locations in the target software domain. The incoming-flow enforcement point (IF-EP) checks the incoming flow of PI data for both the PI creation and update states. The target-software enforcement point (TS-EP) checks what PI data is stored in the domain and how it is consumed, updated, and terminated in the target software domain. The outgoing-flow enforcement point (OF-EP) checks how PI data is transmitted to another software domain.

Target software domain

The scope of the target software domain includes all application programs (Internet browsers, e-mail systems, database applications, domain-specific applications such as order fulfillment, etc.) running on any computing platform (legacy system, workstations, PCs, PDAs, mobile devices, etc.) This enables the PI-LM model to control and monitor PI data in any situation.

ADAPTING THE PI-LM MODEL TO AN IT INFRASTRUCTURE

In this section, we describe how to apply the PI-LM model proposed in the previous section to the IT infrastructure of an enterprise. *Figure 2* shows the system-level relationships among domains, the PI policy enforcers, and the PI management component. The details for the two examples of domains shown in the figure, the Web server domain and the file system domain, are described later. The PI policy-management component provides a centralized management function for associating the policies and the domains. Each PI policy enforcer is deployed to the corresponding target software

domain to check whether the associated PI policies are complied with. The PI policy enforcers send back the results of checking to the PI state-management component, which provides a visualization tool for the CPO.

In adapting the PI-LM model to the existing IT infrastructure, it is important to choose the appropriate target domains of the enterprise. It is also important to implement the PI-PE for each domain. We describe the typical way the model is adapted in the following sections.

Defining domains

A domain may have one or more subdomains. This allows one application to have one or more middleware components, such as databases and legacy systems. The top of Figure 2 shows an example of the PI-LM model adaptation for a Web server domain that includes two subdomains, the file-system domain and the database domain, in its target Web application. For example, the Web server may access the file system to store some data, which may include PI data. At the same time, the file system may be accessed by other applications in an ad hoc manner. By defining the subdomain like this, we can assure that the PI-PE associated with the TS-EP of the primary Web server domain will be able to control the file system domain (as shown on the bottom of Figure 2), because it is defined as a subdomain of the Web server domain.

In a domain, actions or operations for each enforcement point must be taken into consideration. For example, in the IF-EP of the Web server domain, those actions would be file creation and input through Internet browsers from forms which may contain PI data. Thus the PI-PE should be able to handle each action or operation at each enforcement point.

Designing PI policy enforcers

The basic functions of PI-PEs are collecting related information in a domain and detecting points at which some risk exists from the viewpoint of the PI policies, for example, input points at which PI or data that may contain PI comes into a domain.

The characteristics of PI-PEs may vary at each enforcement point. Some may be specific to an enforcement point, and some may be general for the domain. For example, a PI-PE at the IF-EP of a Web server domain, which detects HTML files that contain `FORM` tags to request the input of PI, deeply depends on the HTML format.

In collecting related information in a domain, we focused on checking static information as a first step in developing PI-PEs. It is also necessary to check flow among domains and dynamic changes within domains. These checking functions need to dynamically monitor information flows through networks or inputs using keywords or file accesses. However, if the basic detection functions for PI-PEs are implemented as components of PI-PEs for static information, it is possible to extend those functions to monitor information flows or dynamic changes of content as well.

Deployment in IBM Japan

In adapting the PI-LM model to the IT infrastructure in IBM Japan, we first selected software domains in which PI may be stored that are not strictly controlled and that are widely used among employees. We eliminated specific applications such as an employee salary management system, which obviously handles sensitive personal information, because that kind of system is usually designed in a secure manner with strict access controls, encryption functions, and so on. *Table 1* shows the software domains we selected for our PI-PE deployment in consideration of these issues.

For the PI management function, we implemented a PI-state management database in a Lotus Domino* database, which stored the results of the compliance checks at servers, workstations, and PCs. The association between the software domain and the privacy policy was modifiable by the end users in order to allow flexible administration by each organization. Commercial software, such as Tivoli Security Compliance Manager,[12] was useful in this context for supporting basic functions such as policy management, policy deployment, and report collection from clients to the server.

We developed two PI-PEs: aDesigner, which is a PI-PE for the Web server domain, is described in the section "the aDesigner tool." The Personal Information Detection (PID) tool, which is a PI-PE for the file-system domain and the Lotus Notes* database domain, is described in the section "Personal Information Detection tool."

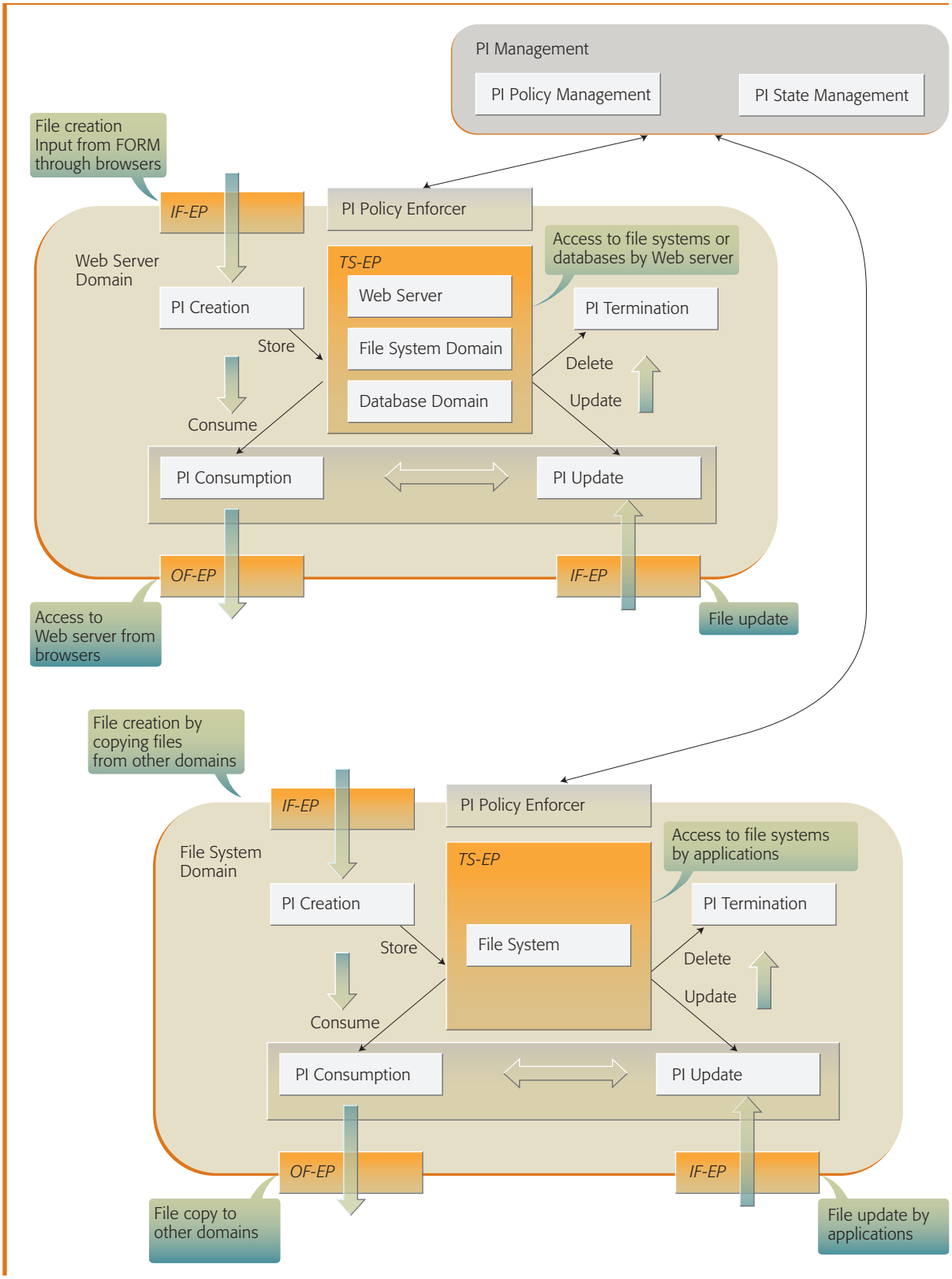

Figure 2
Adaptation of PI-LM model

Table 1 PI Policy Enforcer deployed in IBM Japan by domain

Domain	Subdomain	Enforcement Point	Target Application Data	PI Policy Enforcer
Web server		IF-EP	FORM input	aDesigner
	File system	TS-EP	HTML files in a file system	PID tool for file systems
File systems		TS-EP	Files in a file system	
External devices and media		TS-EP	Files in a file system of external devices and media	
Files in shared file systems		TS-EP	Files in a shared file system	
E-mail		TS-EP	E-mails and attached files in a mail database	PID tool for Lotus Domino databases
Groupware		TS-EP	Documents in a database	

THE ADESIGNER TOOL

This tool is a PI-PE for the Web server domain at the IF-EP in the proposed PI-LM model. When a company collects PI through Web interfaces, it should check for the compliance of such Web pages with its defined Web privacy policy.

The aDesigner tool is a highly accurate compliance checker. (See the subsection "Experimental results" for the accuracy of our checker.) It also provides an easy-to-use GUI (graphical user interface) with features such as the highlighting of places where violations are detected.

The current aDesigner tool evolved from an earlier version[21] that ensured Web pages were accessible to people with visual impairments by simulating the way in which the Web page would appear to this population and by checking for compliance with WCAG (Web Content Accessibility Guidelines). This early tool was enhanced to ensure that enterprise Web sites were in compliance with their defined privacy policy.

A large part of checking for compliance with the accessibility guidelines is based on the structure of the HTML DOM[22] (document object model). For example, the accessibility guidelines WCAG 1.0[23] direct designers to "Provide a text equivalent for every non-text element." For images, this means provide `alt` attributes for all `<img>` tags, which can be verified through DOM manipulation APIs (application programming interfaces). This DOM-based checking can be used to check for compliance with the preponderance of privacy policies.

Figure 3 shows the privacy-checking process of the aDesigner tool. The aDesigner tool performs the checking of a Web site as follows:

1. Crawling through the Web site and collecting Web pages.
2. For each Web page, performing the following subprocesses:
   a. Identifying the language used in the Web page.
   b. Detecting pages which collect PI.
   c. If the page collects PI, checking for compliance with the privacy policy and registering the results in the privacy-compliance-results database.

These steps are described in detail in the following subsections.

Web-site crawling

Currently aDesigner has its own Web crawler. This makes installation of aDesigner easy, but impacts crawling performance. An alternative is to implement aDesigner as a plug-in of IBM WebSphere Information Integrator (WII) OmniFind* Edition V8.4 (OmniFind, for short). See the section "Experimental results" for a discussion of the tool's performance.

Language identification

For each Web page found, the tool identifies the (natural) language used in the targeted Web page from the `lang` attribute of `HTML` tags or from `Content-Language` in `META` tags.

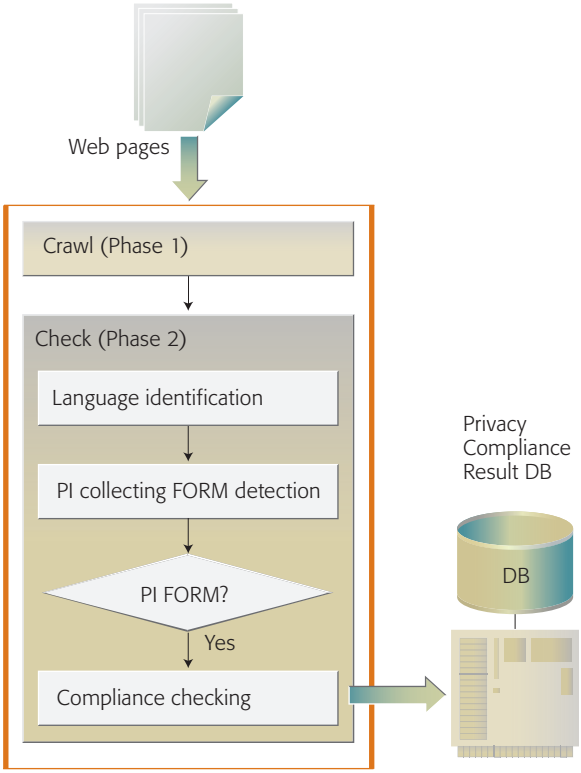

Figure 3
aDesigner Web-site-privacy checking process

Detection of pages that collect PI

The next step is to determine whether each of the collected Web pages are collecting PI. First, Web pages with forms are selected, and then the number of form controls in the page and the labels for them are checked. For example, if a form has fields with labels such as *name, e-mail, phone number, country*, or *region*, the page is identified as a PI-collecting page. On the other hand, if the forms do not collect any PI, or only collect the e-mail address of the customer, which does not identify an individual, the page is not identified as a PI-collecting page. For Web pages written in languages other than English, the judgment is performed in the identified language. See the section "Experimental results" for the accuracy of these judgments.

Compliance checking

PI-collecting pages are checked for compliance with the privacy policy. A large part of the checking is performed based on the HTML DOM. For example, if the policy says "The Web page collecting PI should contain a link to the privacy statement," the checker tries to find an <A> tag with the predefined URL of the statement (defined for each country and language) and with the appropriate link text (the child text element of the <A> tag). A large number of other items in the policy are processed in a similar way. Another type of checking is related to security issues, such as HTTPS (Hypertext Transfer Protocol, Secure) encryption and the form's method type (GET or POST). Checking results are stored in the compliance results database for compliance audits.

Experimental results

Some preliminary experiments were conducted to verify the effectiveness of the approach using aDesigner. These experiments focused on the performance of aDesigner in Web crawling and checking Web pages, and its accuracy in determining whether a page collected PI data.

Crawling and checking time

Our first evaluation was a rough measurement of the crawling performance of OmniFind. The IBM Japan Web site (http://www.ibm.com/jp) was crawled using IBM WebSphere Information Integrator (WII) OmniFind Edition V8.2.2 which was running on an IBM IntelliStation* Z Pro workstation with an Intel Xeon** processor running at 3.06 MHz CPU, 4 GB main memory, and a 120 GB hard disk drive. Using standard settings, approximately 25–30 pages were crawled per second (the performance depended on the condition of the network, etc.). At this rate, crawling all of IBM's external Web pages (about 1.5 million pages) took 17 hours.

The next experiment focused on the time elapsed per page for compliance checking. In this experiment, we used aDesigner for both crawling and checking. We measured the time in two phases—the time to establish a connection with the server and retrieve the page and the time to check for compliance. The crawling and checking target was the IBM Japan Web site, and crawling was performed up to a depth of 1 (i.e., crawling to all child pages from the top page of the Web site, a total of 51 pages). The results show that it takes 195.9 milliseconds to obtain Web pages and 10.4 milliseconds to check them, on average. The slow performance for page retrieval was due to establishing the HTTPS connection (the first phase). The code implementing this function was written in Java. If this function were incorporated in the crawling functionality of OmniFind, the performance would be much improved.

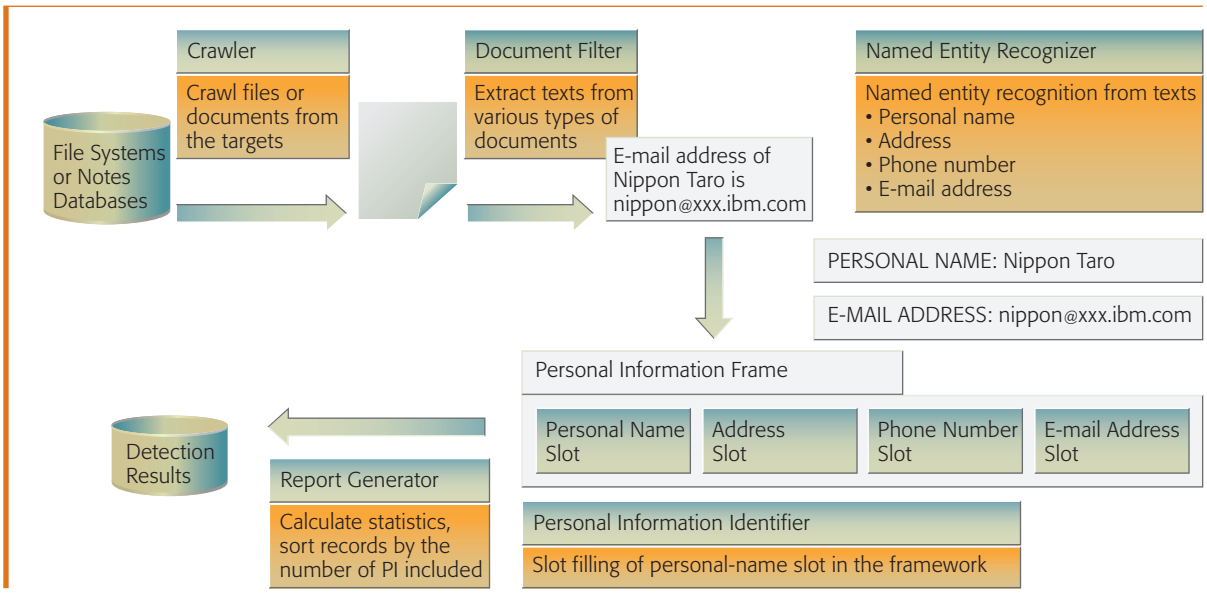

Figure 4
Personal Information Detection tool overview

Accuracy of detecting PI-collecting pages

In detecting PI-collecting pages, two types of errors may occur. One is a false negative, namely when the tool judges the PI-collecting page as a non-collecting one; the other is a false positive, when the tool judges a non-PI collecting page as a collecting one. False negatives have a great impact on the integrity of the entire enforcement system, whereas false positives impact the workload and efficiency of the audit phase.

We conducted an experiment to evaluate the accuracy of aDesigner in detecting PI-collecting pages. The pages in the IBM Japan Web site were crawled and checked by aDesigner. It collected pages up to a depth of 2 (i.e, from the top page to all grandchildren pages, a total of 796 pages). The number of PI-collecting and non-collecting pages were 9 and 87, respectively. We manually checked each page and found that aDesigner had detected all of the PI-collecting pages, that is the false negative ratio was 0 percent; the number of false-positive pages was 3, that is, the false-positive ratio was 0.4 percent, a very high accuracy result.

PERSONAL INFORMATION DETECTION TOOL

The purpose of the PID tool is to support users in detecting PI that is stored in their PCs or other accessible storage. This tool crawls all the files or documents in the specified target file systems or Lotus Notes databases, extracts text from them, and detects PI.

System overview

A system overview of the PID tool is shown in *Figure 4*. The main components of this tool, which are described in the following subsections, are:

1. Crawler
2. Document filter
3. Named entity recognizer
4. PI identifier
5. Report generator

Crawler

A crawler collects files or documents from the specified targets. Two types of crawlers are supported. A file system crawler collects all the files for the specified folders, and a Lotus Notes database crawler collects all the accessible documents and their attached files in the specified databases. A Lotus Notes database crawler is implemented by using a Lotus Notes C++ API.

Document filter

A document filter is used to extract text from various types of files. A document filter called Outside In Technology by Stellent, Inc.[24] was adopted. Sup-

Table 2 File types supported by PID tool

Type	Content Type	File Extensions
Microsoft Word	application/msword	doc
Microsoft Excel**	application/vnd.ms-excel	xls
Microsoft PowerPoint**	application/vnd.ms-powerpoint	ppt
Microsoft Visio**	application/vnd.visio	vsd
Microsoft Project	application/vnd.ms-project	mpp
Lotus 1-2-3	application/vnd.lotus-1-2-3	wj2, wj3, wk3, wk4, 123
Freelance Graphics*	application/vnd.lotus.freelance	prz
Word Pro*	application/vnd.lotus.wordpro	lwp
Adobe PDF	application/pdf	pdf
HTML	text/html	html, htm, shtml, mht
Plain text	text/plain	asc, txt
Rich Text Format	text/rtf	rtf
Comma-separated values	text/csv	csv
Just Systems Ichitaro	application/vnd.justsystem.ichitaro	jaw, jtw, jbw, juw, jfw, jvw, jtd, jtt
Just Systems Matsu	application/vnd.justsystem.matsu	bun
Fujitsu OASYS	application/vnd.fujitsu.oasys	oas, oa2, oa3
XHTML	application/xhtml+xml	xhtml, xht
XML	application/xml	xml, xsl
Log file	application/log	log

ported file types for this tool are shown in *Table 2*. When a file passes through a document filter, a text segment of the file is extracted.

Named entity recognizer

Named entity recognition is a task which extracts elements of named entities, such as names of persons, organizations, locations, and various types of numeric expressions, such as time, date, length, and phone numbers. Named entity recognition is an established research topic[25,26] and is used for various kinds of applications, such as information extraction, information retrieval, question answering, and text mining.

The named entity recognizer used in the PID tool extracts various types of named entities, as shown in *Table 3*. The types shown in bold in the table are used to identify PI.

In named entity recognition, input texts are initially morphologically analyzed into words with their part-of-speech codes. In analyzing a text, rules and dictionaries are used. In the next step (recognition), named entities are recognized from the results of the morphological analysis. In recognizing named entities, dictionaries such as zip-code or area-code dictionaries are used to identify addresses or phone numbers in text. Rules, in which patterns for specified types of named entities are described, are used to identify named entities. For example, a mailing address is recognized by comparison with standard mailing-address patterns.

Named entity recognition is language-dependent because a text first must be analyzed linguistically. Cultural dependencies also play a role in identifying named entities. For example, zip codes or area codes differ in each country. Currently, we have imple-

Table 3 Element types of examples for named entity recognition

Type	Example
Person	Taro Nippon
Address	Minato-ku, Tokyo-to
Zip code	105-0000
Country	Japan
URL	http://www.ibm.com
E-mail address	nippon@xxx.ibm.com
Phone number	(03) 111-1111
Organization	Ministry of Finance
Company	IBM
Bank account number	1000-000-0000000
Credit card number	1111-2222-3333–4444
Date	August 1, 2007
Time	10:30 a.m
Time duration	10 hours
Currency	10,000 yen
Ratio	100%
Ordinal number	1st
Numeral	1
Volume	1 liter
Area	1 square meter
Weight	1 kilogram
Length	1 meter

mented named entity recognition for Japanese. Named entity recognition for English is under development.

The performance of various named entity recognition systems for Japanese were evaluated in the IREX (Information Retrieval and Information Extraction) project.[26] The results indicated an overall f-measure value (i.e., the weighted, harmonic mean of precision and recall) for general-domain newspaper articles as 70, with 84 being the value for the best system evaluated.

Table 4 Precision of named entity recognition

Type	Correct	Total	Precision
Personal name	303	341	88.86
Address	33	49	67.35
Zip code	22	22	100
E-mail address	2	2	100
Phone number	112	115	97.39
Total	472	529	89.22

Table 5 Recall of named entity recognition

Type	Correct	Total	Recall
Personal name	303	322	94.10
Address	33	42	78.57
Zip code	22	22	100
E-mail address	2	2	100
Phone number	112	114	98.25
Total	472	502	94.02

We evaluated our named entity recognition system by using call log data from the IBM PC Call Center only for data types that are used to identify PI. The total number of sentences in the log was 10,000, and its size was 447 KB. The PC used in this evaluation was an IBM Intellistation Z Pro (with an Intel Xeon processor running at 3.06 GHz). Processing time was 5 seconds. The precision for each data type is shown in *Table 4*, and recalls (i.e., correctly detected PI items) are shown in *Table 5*. The resulting f-measure value was 91.56.

Personal information identifier

A "personal information frame" is used to recognize PI. The PI frame is a framework used to represent knowledge. The named slots in this framework represent items of PI, including a personal-name slot, an address slot, a phone-number slot, and an e-mail-address slot. The results of named entity recognition are passed to the PI frame, and the slots are filled with the appropriate values. Each slot has a "lifetime" in numbers of words, which indicates the proximity of this value to other values that this

slot may take. When a slot is filled by a new word, its lifetime counter is reset to a predetermined value (e.g., 30 words). Counters of filled slots are decremented as each word of the named-entity-recognition results is passed to the frame. The frame is refreshed when the count of a slot becomes zero. By clearing the frame when a lifetime becomes zero, values that occur over a limited distance will not be associated as PI that identifies a unique person. When the frame is refreshed, the information in the frame is checked to ensure that it is sufficient to identify a specific person; that is, a personal-name slot and at least one other slot (address, phone number, or e-mail address) is filled.

Report generator

This process generates a report of detection results. These include a summary of the detection task (processing time, total number of scanned files, number of files in which PI is detected, and so on), a list of files with PI, names, and specified keywords, and a list of encrypted files which the tool failed to open.

Process flow in PI detection

Our process to detect files with PI involves three steps: a scan is initiated, detection results are checked, and PI is registered in a corporate PI inventory.

Scan initiation

First, it is necessary to set scan targets for PI detection. There are two types of scan targets, file systems (including local file systems, network file systems, or file systems in external devices and media) and Lotus Notes databases. The user starts the scan by selecting folders in file systems or Lotus Notes databases. Optionally, the user can specify keywords to be detected.

Checking detection results

After the completion of scanning, detection results are shown. In the top page, a summary of detection results is shown, which includes the start and end times of scanning, total number of files and documents in which PI was detected, total number of files and documents in which personal names were detected, and so on.

There are three other views for the detection results. The "PI-detected files" view shows a list of all the files or documents with PI, ordered by the amount of detected PI. The "personal-name-detected files" view shows a list of all the files or documents in which the tool found personal names, but did not find any PI. These files are listed because (in light of the fact that the precision of this tool is not 100 percent) the user may wish to manually check for PI that may have been missed. The "keyword detected files" view shows files containing specified keywords. The tool searches for files that include the specified keywords, such as "confidential," "address list," or specific customer names. In addition to these three views, the tool supplies a list of files and documents that the tool could not open because they were encrypted or password-protected by the function of applications such as Microsoft Office or Windows File System (i.e., NTFS - New Technology File System). These may include important information, for example, customer information. The user may see the contents of the listed files in each window and delete or encrypt them.

Registering PI

After checking the files and documents listed in the detection results, if there are some files with PI that must be kept, the files should be registered manually in a corporate PI inventory. This will be described in the next section.

PI MANAGEMENT—BEST PRACTICES AND EXPERIENCE

IBM Japan manages PI by using the process shown in ***Figure 5***. The privacy policy definition process includes investigating existing laws and privacy policies and defining privacy policies and internal procedures. The defined policies are published internally and learning materials are also published for employees' awareness. Finally, the management process regularly checks for compliance with the policies in each domain and stores the results in an inventory database.

In the Web server domain, aDesigner can be used to check Web pages. Currently, aDesigner is not used in IBM Japan. In the file system, e-mail, and groupware domains, the PID tool can be used to check for files in which PI is included.

Registration in the PI inventory database

The Japanese Ministry of Economy, Trade, and Industry released guidelines for the protection of PI in the area of economy, trade, and industry as a supplement to the Personal Information Protection

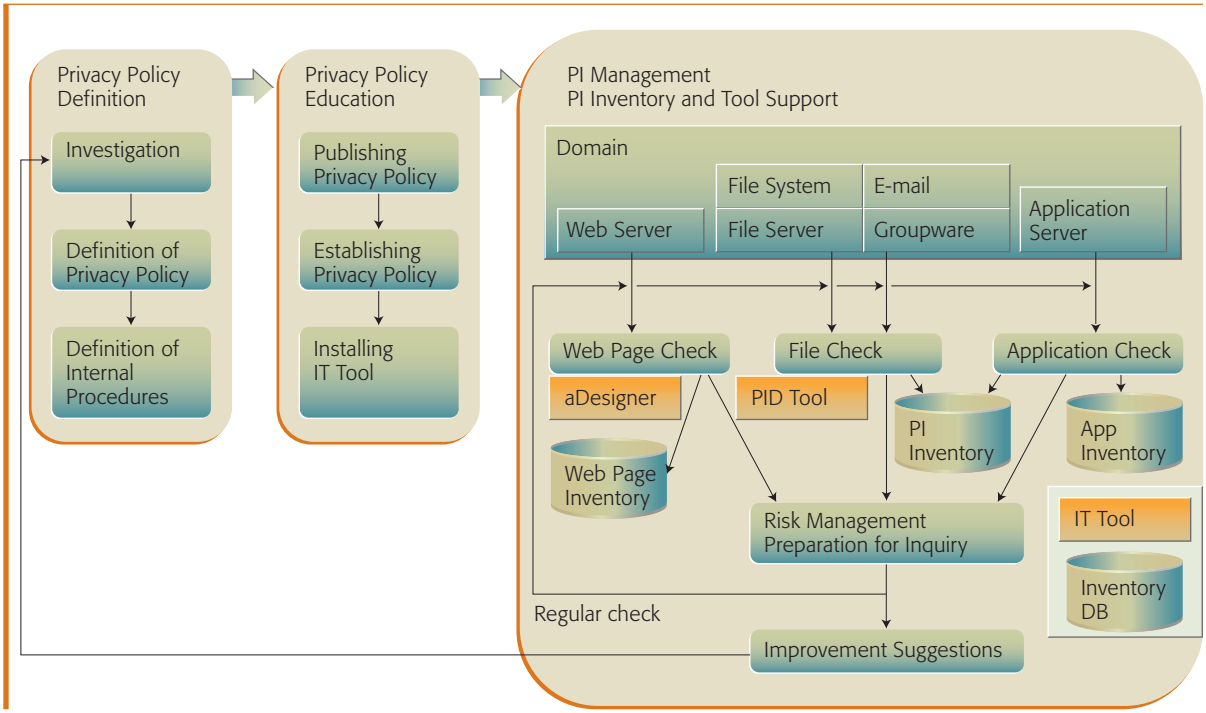

Figure 5
Personal information management process

Act. The guidelines recommend procedures for setting up an administrative system for PI. To comply with the guidelines, IBM Japan established the PI inventory database and requested all employees to register meta-information about PI in the database.

The process of registration in the PI inventory database is as follows. First, employees specify the places where PI may be stored. Second, employees verify that PI exists in those locations. If PI is found, employees categorize it according to risk levels (i.e., the levels of damage or loss to the PI owner in case of PI leakage). Finally, employees register the information in the PI inventory database according to their risk levels. Risk levels are not levels of a data protection scheme, such as encryption. The employees must register the meta-information of the files that contain PI even if the files are encrypted.

IBM Japan assumes that domains where PI is stored are classified into four types of systems: application servers, Lotus Notes databases, file servers, and PCs. A person with responsibility for managing the PI is assigned for each system (see ***Table 6***).

If the data is a PI database, such as a list of names as defined in the guidelines, the employees classify the PI and the risk levels based on the categories defined by the Privacy Office as shown in ***Table 7***. For example, consumers' PI and high-sensitivity information are classified at the high-risk level. As mentioned previously, this classification is independent of both the security level (such as file encryption) and of the location where the file is stored.

The information registered in the PI inventory database needs to be kept up to date. If the PI is deleted from the system, the responsible person inputs the date of its disposal in the registered entry. If the responsible person is absent when the change occurs, e-mail requesting an entry update is sent to the division staff and a manager. If all of the designated people are unavailable, then e-mail is sent to the risk management office.

Employee education

To implement the PI management process, it is important that each employee be aware of the

Table 6 Responsible person for each system

System (Domain)	Responsible Person
Application servers (application server, Web server)	Application owners
Lotus Notes databases (mail, groupware)	Database owners
File servers (files in shared file systems)	Directory owners
PCs and removable media (file systems, external devices, and media)	PC owners or administrators

importance of the privacy policy and of PI management.

Before the enforcement of PIPA in March 2005, IBM Japan released a training course named "Protection of Personal Information is Your Business." In this course, employees learned about PI management and the policies of IBM Japan. In addition to the training, a new course named "Security Training" was released in May 2006. All employees are required to complete these e-learning courses, and the percentage of employee participation is 100 percent for each course.

Statistical results

In this subsection, we present statistics associated with the registration of compliance results and PI characterization in the PI inventory database.

The distribution of personal information by systems is as follows: 69 percent of PI was stored on PCs, 14 percent in Notes databases, 15 percent in application servers, and 2 percent in file servers.

The outer circle in ***Figure 6*** shows the distribution of PI by risk levels. The inner circle shows the distribution by PI category for each risk level. From Figure 6, it can be seen that half of the PI stored is characterized as high risk. Approximately 70 percent of the high-risk PI (35 percent of the total) is highly sensitive employee PI. This is because many managers retain information on their subordinates, such as human resource descriptions or performance information. Figure 6 also shows that the PI of consumers occupies 10 percent of the total, whereas the PI of employees occupies more than 50 percent of the total.

Table 7 PI categories and risk levels

Category	Description	Risk Level
Consumer's personal information	All information about a consumer (e.g., name, address, phone number, birth date, e-mail address, credit card number, purchase history)	High
Highly sensitive business customer information	Private personal information of corporate staff (e.g., birth date, home address, home phone number)	High
Highly sensitive employee information	Private personal information (e.g., information related to human resources, health care, performance, home address, home phone number, bank account, family structure)	High
Business customer information of medium sensitivity	More sensitive information than contact information (excluding highly sensitive information)	Medium
Employee information	More sensitive information than contact information (excluding highly sensitive information)	Medium
Contact information for business customer	Business contact information for customer (e.g., name, e-mail address, divison, company address)	Medium
Contact information for employee	Business contact information of an employee	Low

Consumer information and highly sensitive business customer information constitutes only 13 percent of the PI. In other words, this PI is not diffuse, but concentrated. We believe this is because highly sensitive PI is well-controlled under the privacy policies of IBM Japan.

Normally, the responsible people have to check their systems manually to find PI. Such manual checking may take an extreme amount of time and may nevertheless miss some PI. The PID tool can scan the files in a PC or a file server, and the documents and attachments of Lotus Notes databases. The PID tool covers 85 percent of all systems (69 percent for PCs, 14 percent for Notes databases, and 2 percent for file servers). In addition, ***Table 8***, which shows the distribution of PI by risk levels and systems, indicates that the PID tool covers over 90 percent of high-risk PI.

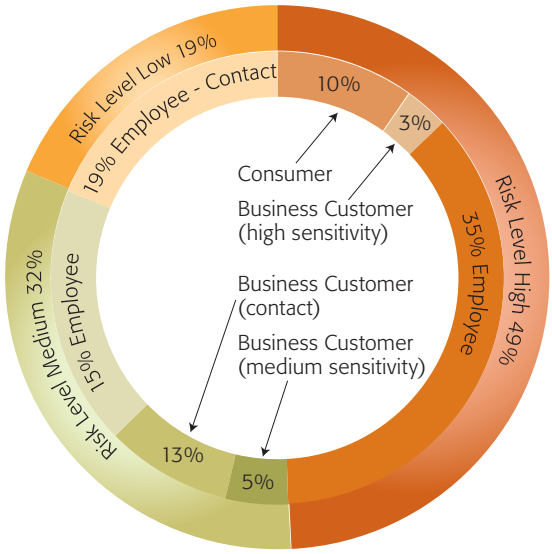

Figure 6
Distribution of personal information

DISCUSSION

In this section, we discuss several issues and possible enhancements related to PI privacy management.

On-the-fly compliance check

The current implementation of our PID tool assumes the regular use of the tool, for example, once a week or once a month, in order to update the status of PI. This use scenario is appropriate when the target IT environment is a general domain which is used by well-educated employees in their daily work. In contrast, if the target IT environment deals with highly sensitive PI, such as credit card numbers in some mission-critical applications, or the employees are not educated in privacy policies, stricter and more fine-grained enforcement mechanisms should be applied. In such an environment, the PI enforcer should provide an on-the-fly compliance checking mechanism to make sure that PI cannot be stored in any local file systems or removable media. Such an on-the-fly PI enforcer observes any changes in PI data and monitors any flow of PI data at the IF-EP or the OF-EP of our PI-LM model. The detection logic we developed for the aDesigner tool and the PID tool can also be incorporated with a commercial operation monitoring agent, which enables the on-the-fly compliance checking mechanism.

Application of the PI-LM model to a non-IT environment

We applied the PL-LM model only to the IT infrastructure to develop automated compliance-checking tools. However, this model can also be applied to non-IT environments by use of manual processes to find policy enforcement points and related software domains, to propose necessary actions, and to define the PI-PEs. The output of this application would be a set of required procedures or assignments for persons in charge, designed to reduce risks in PI management.

Data-mining function for PI-state management

In this paper, we have not focused on functions needed for centralized PI-state management in which the "plan-do-check-action" cycle helps the chief privacy officer manage the PI-related risks in the company. We believe that data-mining technologies would be useful in calculating risks based on the PI status collected from the PI-PEs.

Table 8 PI distribution by system for each risk level (percentage)

	PCs	Notes DBs	File Servers	Application Servers	Total
High	74.6	16.1	2.1	7.2	100.0
Medium	65.8	11.5	2.1	20.6	100.0
Low	56.7	14.5	1.0	27.8	100.0

CONCLUSION

We have presented a holistic PI life-cycle management model called the PI-LM model and described two PI-PE tools, the aDesigner tool and the PID tool, which are positioned at typical policy enforcement points to enable a chief privacy officer to grasp the status of the PI in the company.

We showed that aDesigner can examine a large volume of HTML pages very quickly. The PID tool has been deployed in all of the PCs owned by IBM Japan employees, and the reports from the tool are stored in the centralized PI inventory database, which helps the Chief Privacy Officer to grasp the current status of PI in IBM Japan.

We presented a description of the best practices used by IBM Japan and the statistical results about how PI is stored and managed in the company. The statistics show that high-sensitivity information is stored in many PCs rather than on servers, and the PID tool has been effective in identifying such situations without requiring excessive effort by the employees. We are sure that the deployment of the two PI-PE tools will reduce the cost of PI management in IBM Japan and also reduce the risk of unexpected PI leakage incidents by enabling countermeasures to be taken based on current PI status. We plan to extend the current PID tool to support multiple languages, including English.

ACKNOWLEDGMENTS

We would like to thank Naoki Kobayakawa for providing data obtained through practical experience in PI management in IBM Japan. We are also grateful to Satoshi Hada and Naishin Seki who developed the original PID tool.

CITED REFERENCES

1. "Financial Data Lost by Bank of America," *The Boston Globe* (February 26, 2005), http://www.boston.com/business/articles/2005/02/26/financial_data_lost_by_bank_of_america/.
2. *Identity Theft Survey Report*, Federal Trade Commission (September 2003), http://www.ftc.gov/os/2003/09/synovatereport.pdf.
3. *Act on the Protection of Personal Information (kojin jouhou no hogo ni kansuru houritsu)*, Cabinet Office, Government of Japan (2003), http://www5.cao.go.jp/seikatsu/kojin/foreign/act.pdf.
4. "The Result of Survey on Communication Usage Trends in 2005" (in Japanese), Ministry of Internal Affairs and Communications, http://www.soumu.go.jp/s-news/2006/060519_1.html.
5. IBM Privacy Practices on the Web, IBM Corporation, http://www.ibm.com/privacy/us/.
6. L. Cranor, M. Langheinrich, M. Marchiori, M. Presler-Marshall, and J. Reagle, *The Platform for Privacy Preferences 1.0 (P3P1.0) Specification*, Worldwide Web Consortium (April 2002), http://www.w3.org/TR/P3P/.
7. *eXtensible Access Control Markup Language (XACML) Version 2.0*, T. Moses, Editor, OASIS Standard, Oasis Open Consortium (February 1, 2005), http://docs.oasis-open.org/xacml/2.0/access_control-xacml-2.0-core-spec-os.pdf.
8. M. Backes, B. Pfitzmann, and M. Schunter, "A Toolkit for Managing Enterprise Privacy Policies," *Proceedings of the 8th European Symposium on Research in Computer Security (ESORICS), Lecture Notes in Computer Science* **2808**, Springer-Verlag, Berlin (2003), pp. 162–180.
9. P. Ashley, C. Powers, and M. Schunter, "From Privacy Promises to Privacy Management—A New Approach for Enforcing Privacy Throughout an Enterprise," *Proceedings of the ACM New Security Paradigms Workshop*, ACM Press, New York (2002), pp. 43–50.
10. C. Giblin, A. Y. Liu, S. Müller, B. Pfitzmann, and X. Zhou, "Regulations Expressed As Logical Models (REALM)," *Proceedings of the 18th Annual Conference on Legal Knowledge and Information Systems (JURIX 2005)*, IOS Press, Amsterdam (2005), pp. 37–48.
11. M. Numao, Y. Watanabe, M. Yuriyama, T. Yoshizawa, and C. Powers, *Application Privacy Monitoring for JDBC*, IBM AlphaWorks (2004), http://www.alphaworks.ibm.com/tech/apm4jdbc (2004).
12. *IBM Tivoli Security Compliance Manager*, IBM Corporation (2006), http://www-306.ibm.com/software/tivoli/products/security-compliance-mgr/.
13. P-Pointer (in Japanese), KLab Security, Inc., http://www.klabsecurity.com/product/p-pointer/index.html.
14. Sumizumi-kun (in Japanese), Mitsubishi Space Software Co., Ltd., http://www.mss.co.jp/businesfield/security/sumizumi/index.html.
15. eX PDS (Privacy Document Searcher) (in Japanese), Quality Corporation, http://www.quality.co.jp/products/eXPDS/.
16. Kenshutsu Meijin (in Japanese), Toyama Fujitsu Ltd., http://jp.fujitsu.com/group/tfl/services/kensyutsu/index.html.
17. K. Yasu, Y. Akahane, M. Ozaki, K. Semoto, and R. Sasaki, "Evaluation of Check System for Improper Sending of Personal Information in Encrypted Mail System," *IPSJ (Information Processing Society of Japan) Journal* **46**, No. 8 (2005), pp. 1976–1983.
18. H. Sakaki, K. Yanoo, R. Ogawa, and I. Hosomi, *An Information Leakage Risk Evaluation Method Based on Sensitive Document Detection and Security Configuration Validation*, IEICE (The Institute of Electronics, Informa-

tion, and Communications Engineers) Technical Report **105**, No. 395 (2005), pp. 15–22.

19. WebXM Overview, Watchfire Corporation, http://www.watchfire.com/products/webxm/.
20. H. Ryan, P. Spyns, P. De Leenheer, and R. Leary, "Ontology-Based Platform for Trusted Regulatory Compliance Services," *Proceedings of the OTM Confederated International Workshops—On The Move to Meaningful Internet Systems (OTM 2003), Lecture Notes in Computer Science* **2889**, Springer, Berlin (2003), pp. 675–689.
21. H. Takagi, C. Asakawa, K. Fukuda, and J. Maeda, "Accessibility Designer: Visualizing Usability for the Blind," *Proceedings of the 6th International ACM SIGACCESS Conference on Computers and Accessibility (Assets '04)*, ACM Press, New York (2004), pp. 177–184.
22. *Document Object Model*, W3C Architecture Domain, Worldwide Web Consortium (2005), http://www.w3.org/DOM/.
23. *Web Content Accessibility Guidelines 1.0*, W3C Recommendation, Worldwide Web Consortium (1999), http://www.w3.org/TR/WCAG10/.
24. Outside In Technology, Stellent, Inc., http://www.stellent.com/en/products/outside_in/index.htm.
25. Introduction to Information Extraction, National Institute of Standards and Technology (2005), http://www-nlpir.nist.gov/related_projects/muc/index.html.
26. S. Sekine and Y. Eriguchi, "Japanese Named Entity Extraction Evaluation—Analysis of Results," *Proceedings of the 18th Conference on Computational Linguistics (Coling 2000), Vol. 2*, Association for Computational Linguistics, Morristown, New Jersey (2000), pp. 314–321.

Accepted for publication November 9, 2006.
Published online April 18, 2007.

Michiharu Kudo
IBM Research Division, Tokyo Research Laboratory, 1623-14, Shimotsuruma, Yamato-shi, Kanagawa-ken, 242-8502, Japan (kudo@jp.ibm.com). Dr. Kudo received B.E., M.E., and D.E degrees from the University of Tokyo in 1986, 1988, and 2002. He joined the IBM Tokyo Research Laboratory in 1988 and has worked on information security research. He is currently a manager of the Security and Privacy group at the IBM Tokyo Research Laboratory. Dr. Kudo has been a part-time lecturer at the Tokyo Institute of Technology since 2000. He is one of the founders of the XACML Technical Committee of the OASIS standardization body. His research interests include XML security, privacy protection, and access control technology.

Yoshio Araki
IBM Japan, Ltd., 2-12, Roppongi 3-chome, Minato-ku, Tokyo 106-8711, Japan (araki@jp.ibm.com). Mr. Araki has been the Chief Privacy Officer of IBM Japan since October 2001. He received an M.S. degree in communication engineering from Osaka University and subsequently joined IBM Japan as a development engineer at the Fujisawa Development Laboratory. He worked on various products, including imaging devices, communication devices, host-to-PC communication software, and mobile network software. He moved to product planning in 1989 and did market planning in Raleigh, North Carolina, and Boca Raton, Florida from May 1990 to December 1992. After several management positions in development, he served as Asia Pacific marketing manager in mobile computing in 1997. He spent 1998 and 1999 in Strategy and Business Development of IBM's Asia Pacific Technical Operations to incubate security-related projects. From January 2000, he has worked as a marketing manager in Network and Systems Development.

Hiroshi Nomiyama
IBM Research Division, Tokyo Research Laboratory, 1623-14, Shimotsuruma, Yamato-shi, Kanagawa-ken, 242-8502, Japan (nomiyama@jp.ibm.com). Mr. Nomiyama joined the IBM Tokyo Research Laboratory in 1985 after receiving an M.S. degree from Kyushu University. He has been involved in Japanese-to-English machine translation, information retrieval, and text-mining projects. His research interests include natural language processing, text mining, and information visualization.

Shin Saito
IBM Research Division, Tokyo Research Laboratory, 1623-14, Shimotsuruma, Yamato-shi, Kanagawa-ken, 242-8502, Japan (shinsa@jp.ibm.com). Mr. Saito received an M.S. degree in information science from the University of Tokyo in 2001 before joining the IBM Tokyo Research Laboratory. His research interests include Web accessibility and usability and static analysis of mark-up and programming languages.

Yukihiko Sohda
IBM Research Division, Tokyo Research Laboratory, 1623-14, Shimotsuruma, Yamato-shi, Kanagawa-ken, 242-8502, Japan (sohda@jp.ibm.com). Dr. Sohda received B.S., M.S., and Ph.D. degrees from the Tokyo Institute of Technology in 1998, 2000, and 2003. Since joining the IBM Tokyo Research Laboratory, he has worked on Web Services caching. His research interests include high-performance parallel architectures, privacy management, and compliance management. ■