Information Commons: Concept of Disclosing Public Information in an Era of Personal Privacy and Information Protection

● Toshihiro Enami

The Internet has been eliminating the asymmetry of information between governments and citizens and has been helping to foster public awareness in Japanese society. However, a lot of information that should be public is being concealed in the name of personal data protection. To promote the disclosure and sharing of such information, we should create a world in which public information can be shared among every citizen, while personal data is protected at the same time. That is, we should embrace the concept of "Information Commons." This paper introduces the concept of Information Commons and the technologies for realizing it, focusing on the significant social impact these technologies would have in the applied fields.

1. Introduction

The amount of information circulating through society has been increasing with the spread of the Internet. This is due not only to an increase in communication among people, but also to governments disclosing public information on their Web sites. In this way, the free distribution of public information has contributed to the advance of a democratic society, as the fostering of public-issues awareness has made the government more familiar to citizens.

However, a phenomenon that reverses the tide of democracy is also occurring. The advance of an information society has brought with it a heightened sensitivity to the frequent abuse of personal data, and regulatory systems for protecting this information have been instituted. As a result, public information is being concealed in the name of personal data protection, and the development of a democratic society has been hindered and in some cases even reversed.

Technology that can solve the conflict between protecting personal data and disclosing public information is necessary to avoid regressing back to the dark ages when information was controlled by a select few and the general population was without a voice. This paper proposes the concept of "Information Commons," which can solve one of the biggest challenges of an information society, and describes the technologies for realizing it.

2. Social issues brought on by awareness of personal data protection

In the last decade, the emergence of the information society has been greatly facilitated by the spread of the Internet. These days, people can exchange text messages and search for information on their mobile phones anywhere. At train stations, it is now possible to board a train or shop using electronic money in a smart card. A ubiquitous digital society has emerged in which IT services that make life more convenient are

FUJITSU Sci. Tech. J., **43**,2,p.159-167(April 2007)

159

available anytime and anywhere.

On the other hand, personal data has been recognized as a valuable resource in this society, and the selling of personal data has become a business. This trend has made people more sensitive to personal data abuse, while damage compensation suits have been appearing more frequently in the courts.

For example, in May 1999, personal data of citizens of Uji city was leaked from the Uji city office. A part-time worker hired by a company to do data-input work for the city sold 210 000 pieces of personal data concerning Uji citizens to a name-list business company. A suit followed, and the court ruled that the city government pay ¥10 000 to each claimant. The fact that this incident occurred at a town office and that a law concerning government protection of personal data protection called the Law of Personal Data Protection of Government-Run Computer Systems had already come into effect in 1988 was a significant shock to society.

In another case, roughly 4 500 000 pieces of customer data were leaked from Yahoo!BB in February 2004. This incident was significant enough to pose a threat to the company as a whole. The general public was surprised at not only the amount of leaked data but also at the fact that it was used to threaten the company. Furthermore, the compensatory payment of more than 2 billion yen in ¥500 cash vouchers to customers was a grim eye-opener for the world of corporate management.

In April 2005, a law concerning personal data protection applied to private companies took full effect. The law stipulates that up to six months of incarceration or a fine of less than ¥300 000 is to be imposed on persons who fail to obey an order of a state minister. However, for both government administrations and private companies, the loss of social trust and legal/financial damages are considered more serious than criminal punishment.

As a result, government administrations and private companies, fearing the loss of social trust and legal/financial damages, are beginning to avoid digitizing data, promoting IT, and/or disclosing information. This trend has led to a fear of using IT and the concealment of information, which are connected with the following issues regarding over-protection of personal data as reported by the *Yomiuri Shimbun* newspaper on August 3, 2005.

1) Caseworkers cannot observe or follow-up on cases of abuse among families because relevant information is not provided by the government.
2) The legitimacy of cases of *amakudari* (the practice of giving high-paying corporate jobs to retiring government officials) cannot be checked because the *amakudari* lists are not disclosed.
3) The competency of teachers cannot be checked because information concerning poor teacher performance is not disclosed.
4) The government does not disclose information about workers' compensation claims of asbestos accidents.
5) Hospitals do not provide the police with information about the status of patients delivered to hospitals after accidents.
6) Schools do not create class lists or emergency network lists.

3. Developing Information Commons

Protecting personal data is therefore very important; but at the same time, overprotection produces serious social issues. The mentality of avoiding information disclosure and the distribution of information is against the concept of a society opened up by the free distribution of information over the Internet. In the past decade, the Internet has rapidly advanced the sharing of information in society, and in doing so it has contributed to the progression of democracy by eliminating the asymmetry of information between governments and citizens. Information

distribution is essential in establishing a democracy, as reflected in Ralph Nader's[note] assertion that, "information is the currency of democracy." If information is concealed in the name of personal data protection, our society will return to an undemocratic dark age in which there is little information distribution.

I propose the concept of Information Commons, which, like common land and common ground, refers to public property. It indicates a world in which members of society share public information while simultaneously protecting personal data, and the technological architecture used to realize this world is called Information Commons technology. Again, I emphasize the following.

It is true that the Internet has helped eliminate the asymmetry of information between governments and citizens and has fostered an awareness of public issues. However, public information is being concealed in the name of personal data protection. Public information should be disclosed in a positive way and shared among members of society.

In order to foster public awareness and build an even better society, we must first create a world in which everybody can access public information while protecting personal data. Information Commons technology can contribute to the realization of this goal.

The main points of Information Commons technology are that it can transform personal data into a privacy-protected form without reducing the original value of the information and that it can guarantee public information is shared within society and is available without charge. In other words, it can automatically conceal personal data or anonymize it by replacing it with placeholder names (e.g., John Doe) while assuring the integrity of the digital information, and in this way it allows public information to be widely shared with a sense of security. Information

---

note)   An American consumer activist.

Commons technology is comprised of the following three technologies, all developed by Fujitsu Laboratories Ltd.
1)   Personal data concealment technology
2)   Logical structure recognition technology
3)   Original Information Assurance technology

4. Information Commons technology

Information Commons technology can automatically conceal or anonymize personal data and assure the integrity of digital information. The deletion function is performed using:
- Personal Data Concealment technology, which accurately recognizes personal data in text data and deletes it, and
- Logical structure recognition technology, which accurately recognizes personal data in image data and deletes it.

The integrity assurance is done using:
- Original Information Assurance technology, which assures all the data other than the concealed and anonymized data is the same as in the original and therefore protects against data corruption and also data falsification.

4.1 Personal data concealment technology[1]

Text data might contain names, birth dates, addresses, and other information that identifies a person. To protect personal privacy, this information must be recognized and transformed into an appropriately different form. Fujitsu Laboratories Ltd. has developed the personal data concealment technology, which can automatically recognize personal data such as names and birth dates in text data and conceal or anonymize this information by using a method developed by Fujitsu called Named Entity Identifying Technology by Learning with Teacher (**Figure 1**).

For example, courts decisions are not generally disclosed because they include personal data, but this technology can transform the data

FUJITSU Sci. Tech. J., **43**,2,(April 2007)

161

Data learning

<person> 松井 </person> 部長は
<person> 岩倉 </person> 君の上司
<person> 岩倉 </person> 君は
<org> 知能研 <org> に所属。

↓ Learning the rule for
extracting named entities

Before of 部長 is <person>
Word including 研 could be <org>

Learned rule for
extracting named entities

Learning phase

Input text

山田部長は田中君の友達。
太郎君は理研の社員。

→ Extracting named entities

山田部長は田中君の友達。
太郎君は理研の社員。

Anonymizing

鈴木（仮）部長は斉藤（仮）君の友達。
佐藤（仮）君は山田商事（仮）の社員。

Extracting and anonymizing phase

(a)

Data learning

The manager
<person>Matsui</person> is a boss of
Mr. <person>Iwakura</person>.
Mr. <person>Iwakura</person>
belongs to <org>the knowledge
laboratory<org>

↓ Learning the rule for
extracting named entities

The next of 'manager' is <person>
Word including 'laboratory' could
be <org>

Learned rule for
extracting named entities

Learning phase

Input text

The manager Yamada is
a friend of Mr. Tanaka.
Taro is an employee of Riken.

→ Extracting named entities

The manager Yamada is a
friend of Mr. Tanaka.
Taro is an employee of Riken.

Anonymizing

The manager Suzuki (anonymized)
is a friend of Mr. Saito (anonymized).
Sato (anonymized) is an employee of
Yamada Shoji (anonymized).
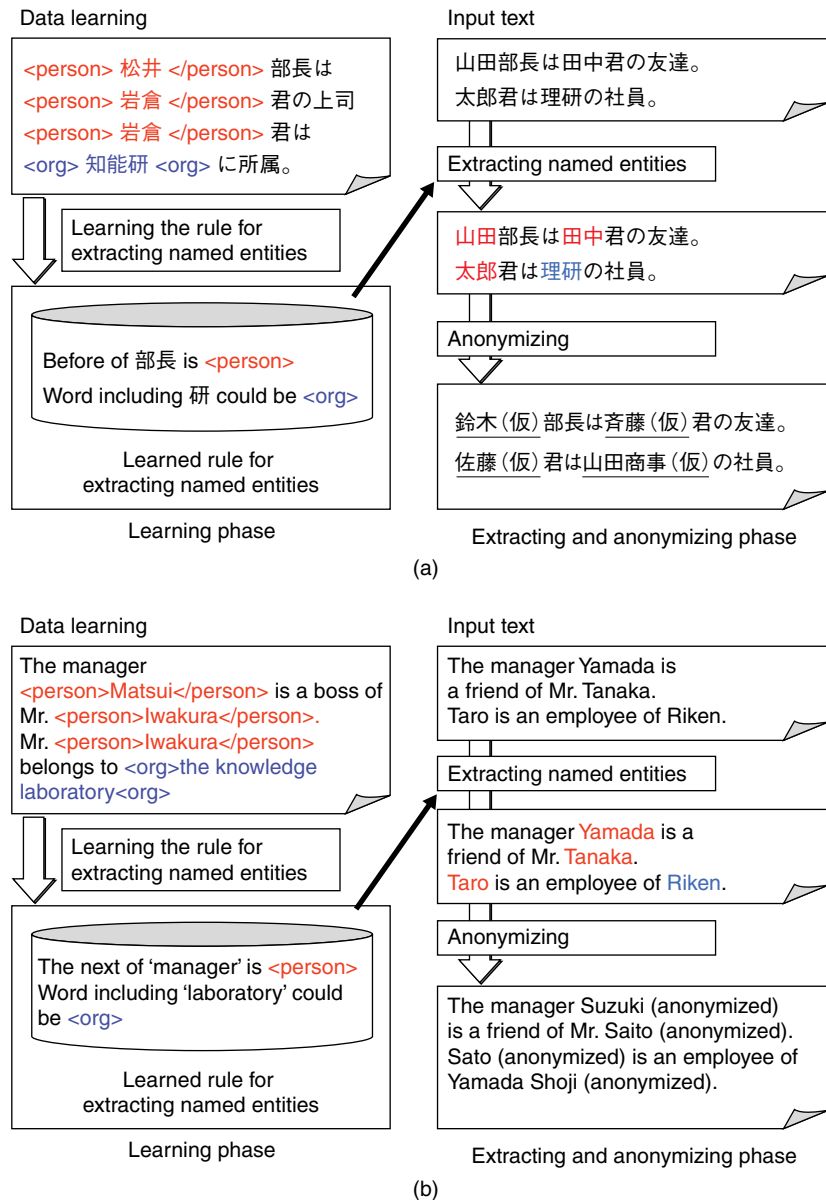
Extracting and anonymizing phase

(b)

Figure 1
Personal data concealment technology. (a) is original, and the Japanese words
on (a) are translated into English on (b) to help readers to understand.

so the decision can be safely disclosed. Further-more, replacing person-identifying data with, for example, "*****," makes the context difficult to understand, but making it anonymous does not. The following is an example in Japanese of a sim-ulated court decision with text data that has been transformed using this technology. To make the text disclosable to anyone, the underlined person-identifying words have been anonymized and some of the text has been replaced with " 某 ".

（被告人の身上経歴）
　被告人は、昭和五十年十一月二十日（仮）
埼玉県あさひ市緑町（仮）において、父義男
（仮）、母良子（仮）の長男として生まれ、同市
立の小学校、中学校に通った。その間被告人が
小学二年生の時に、父親が勤務していた会社が
倒産し、当時係長であった父親は失業した。父
義男は、毎日酒びたりの日々を送り、・・・・・
　（中略）　　・・・・・・
　被告人は、平成三年四月に都内の都立某高等学
校に入学し、平成六年三月に同高等学校を卒業し

> た。・・・（中略）・・・食事も母親に自室に運ばせ、いわゆる引きこもりの状態が続いていた。翌年の平成七年四月、私立某大学理工学部に入学した。

"仮" means the word is anonymized, and "某" means "anonymous."

A translation of this Japanese text is given below. The underlined words are anonymized forms of the original data.

(Profile of the defendant)

The defendant was born as the first son on <u>November 20, 1975 (anonymized)</u> at <u>Midori-ward, Asahi city, Saitama prefecture (anonymized)</u> of his father <u>Yoshio (anonymized)</u> and mother <u>Yoshiko (anonymized)</u>. He went to public elementary school and junior high school. During those days, in the second grade at his elementary school, his father lost his job as an assistant manager because of the collapse of his company. Yoshio became an alcoholic. ……… (omitted) ……….

The defendant entered <u>a Tokyo Metropolis public high school</u> in April 1991 and graduated in March 1994. ……… (omitted)……… . He made his mother take his meals to his room and lived in his room all day, demonstrating acute social withdrawal. On April of 1992 the next year, he entered the department of science and engineering of <u>a private university</u>.

Therefore, we can transform text data into a publishable form that anyone can understand by anonymizing names, birth dates, addresses, organization names and other person-identifying information. Of course, not all data can be transformed completely and automatically, but recognition accuracy can be improved with more research and the amount of manual labor needed to anonymize information can be decreased.

4.2 Logical structure recognition technology[2]

Personal data concealment technology can deal only with text data, but there are also cases where personal data is present in image data. This type of personal data can be detected using Fujitsu Laboratories Ltd.'s logical structure recognition technology (**Figure 2**).

This technology, which can anonymize personal data more widely and accurately when used in conjunction with personal data concealment technology, can do the following.

1) Detect personal and confidential data from a variety of document data, including images.
2) Detect patterns in personal and confidential data from not only text data but also spreadsheet and image data.
3) Detect specified personal data by using models that can identify personal and confidential data patterns.

4.3 Original Information Assurance technology[3]-[5]

When handling public information, it is important that its integrity is preserved. If public information is anonymized, it must be proved that it is identical to the original version except for the anonymized parts.

The Partial Integrity Assurance Technology (PIAT) for signature, which was developed exclusively by Fujitsu Laboratories Ltd., is an important Original Information Assurance technology for confirming the originality of information to third parties (**Figure 3**). PIAT does the following.

1) Confirms that all of the non-anonymized information that the third party can see in the document is the same as in the original version.
2) Detects and identifies falsified parts, including personal data that has been illegally added to replace legally deleted personal data.
3) Identifies both the creator and editor of information (when, who, which parts, and how).
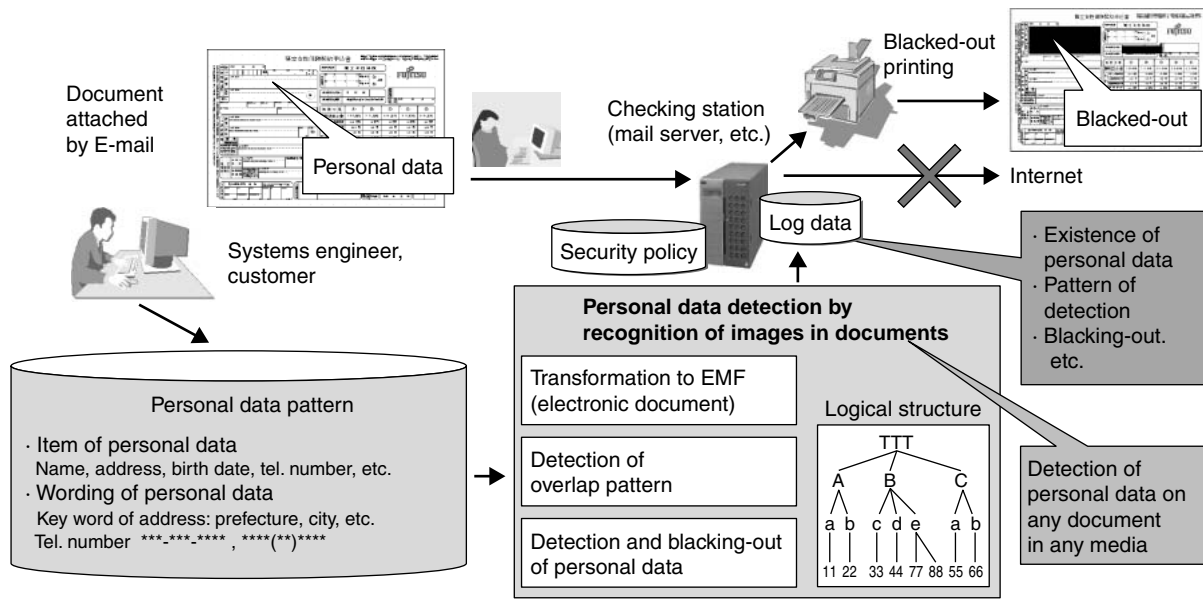
FUJITSU Sci. Tech. J., **43**,2,(April 2007)

163

Figure 2
Logical structure recognition technology.
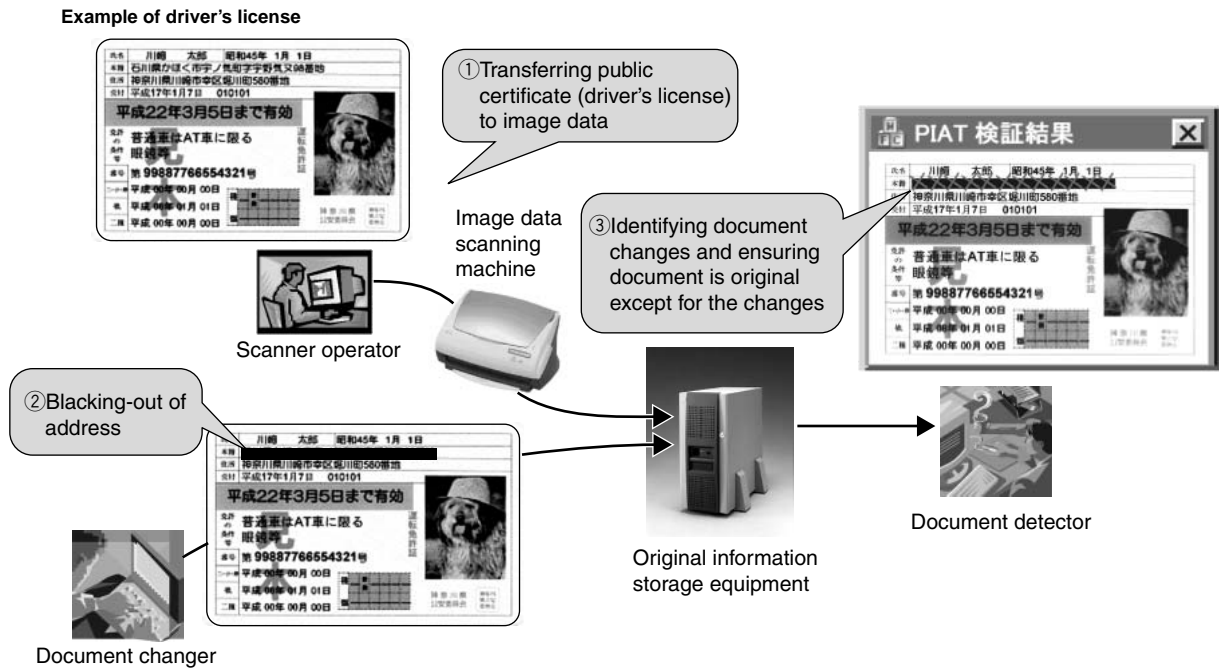

Figure 3
Original Information Assurance technology.

5. Application areas of Information Commons technology

I will now turn to the areas that Information Commons technology can be applied to and the types of current problems that it can solve. The analysis is divided into the following eight sub-sections.

5.1 Disclosure of court decision documents

In May 2003, the public aspects of court rulings were discussed at the 6th SHIP project international symposium in Tokyo. Chris Puplic, President of the Anti-Discrimination Board of New South Wales, raised the issue of dealing with personal data in verdict-related documents. She first asserted that these documents must be shared within society. However, if this information is posted on the Internet, there is the possibility that personal data would be collected and used to create personal profiles. Therefore, she continued, we are faced with the issue of not being able to share court ruling documents on the Internet because of the possibility of personal privacy infringement.

So far, this issue has not garnered much serious attention in Japan, but it is a fact that court ruling documents are not posted on the Internet. In the case of important court decisions, legal publishers anonymize the documents manually and then publish them in magazines. With our technology, court rulings can be anonymized automatically and courts can provide relevant databases on the Internet.

A citizen judge system is scheduled to be introduced in Japan by 2009, which means that ordinary citizens will become involved in law cases. It is therefore very important that court ruling documents and legal information in general are provided to the general public.

5.2 Avoiding infringement of personal privacy through past newspaper articles

In general, newspaper articles are public information. However, past newspaper articles stored in databases are becoming the source of personal privacy infringements. For example, an ex-convict is discriminated against as a criminal because of a past newspaper article.

The current situation is that newspaper companies will anonymize past articles only in response to individual complaints, and newspaper associations have not created any standards with regards to this issue. The text of newspaper articles is formulaic, so technology can be easily applied to it. Therefore, if standards were set concerning the type of information to be anonymized, for example, the date and type of a crime, this technology could be used to assure that past newspaper articles stored in databases would not lead to breaches of personal privacy.

5.3 Disclosure of government documents

Central and local governments must follow laws and ordinances to disclose documents if requested by citizens. However, documents are only released if all personal data has been blacked-out. This can take a considerable amount of manual labor, and in some cases disclosure requests are denied because the document contains too much personal data.

This technology can reduce the work of officers who are involved in disclosure requests and also prevent government documents from being concealed just because they contain personal data. In this way, the technology can support democratic systems that have been weakened by the issue of personal privacy.

5.4 Disclosure of academic data

How to handle data that is used as the basis for academic discussion is another issue. Particularly, in the medical field, doctors often handle private information and must anonymize this data

manually in order to complete their papers or make presentations at academic conferences. This work is cumbersome, and the possibility of disclosing personal data by mistake can lead to significant problems.

This technology can reduce the workload for researchers who deal with personal data, and by checking personal data automatically, it can also help prevent the erroneous release of personal data.

5.5 Producing test data when developing computer systems

As we have seen, this technology can be effectively applied to the government sector, but it can also be effective in the private sector.

For example, real-world data is often used in the final testing of a new computer system. However, it is difficult for system developers to obtain the appropriate data when that data includes personal data. In these cases, system developers are forced to use simulated data, which provides a much lower degree of testing accuracy than real-world data.

Our technology can eliminate this problem. By using our technology, an ordering party can create anonymized data from real-world data and provide it to system developers, who can then run tests with a comparable degree of accuracy to tests using the original data.

5.6 Analyzing data

Recently, there has been an increase in the number of companies that analyze customers' data, recognize consumer trends, and use software for improving sales through the evolving technology of data mining. These companies rarely outsource their data mining work to other companies or even distribute it among their own business sections because of personal data protection issues.

However, information such as names, dates of births, and other information that can identify, for example, a person, is irrelevant in this type of data analysis. Therefore, the data can be anonymized using our technology and then freely distributed and outsourced, and the consequent analyses results would contribute to increased sales.

5.7 Sharing information

In many companies, information is shared for effective management. Instead of starting from scratch, proposals, estimate sheets, and other materials are obtained from other business sections with similar documents. However, when using a proposal document for company A to make a similar proposal document for company B, the document might mistakenly include words and phrases about company A by mistake. If such a document were to be released to the outside world, the credibility of the company that made the mistake could be severely damaged.

Our technology makes it possible to check for words and phrases related to company A in a similar document made for B company, and in this way can help companies promote the effective flow of information.

5.8 Managing personal data

The law concerning personal data protection stipulates that companies must handle personal data in an appropriate manner. However, these days companies distribute so many PCs to their employees that their servers contain so much data it is difficult to determine the locations of all the personal data within the company.

By installing our software on each PC and server in a company, it becomes possible to ascertain the location and nature of all personal data within the company and then implement appropriate personal data protection measures.

6. Conclusion

An increasingly important challenge for the future is to further cultivate the spread of democracy that has evolved with the release of information through the Internet, while simulta-

neously protecting personal data and privacy more appropriately.

We believe that the concepts and technologies presented in this paper offer appropriate solutions for these social issues. We expect that our technology can be applied to a variety of fields to achieve Information Commons; that is, a world in which the members of society can share information that can safely be made public while protecting personal data.

References

1) T. Iwakura, S. Okamoto, and K. Matsui: Assistant Tool for Concealing Personal Information in Text. *FUJITSU Sci. Tech. J.,* **43**, 2, p.212-219(2007).
2) A. Minagawa, Y. Fujii, H. Takebe, and K. Fujimoto: A Method of Logical Structure Analysis for Form Images with Various Layouts by Belief Propagation. PRMU, Oct. 2006.
3) T. Yoshioka and M.Takenaka: The proposal of Partial Original Information Assurance technology for distribution and modifying digital data. 3rd Information Science Technology Forum (FIT2004), M-066, 2004, p.231-232.
4) T. Izu, N. Kanaya, M. Takenaka, and T. Yoshioka: PIATS: A Partially Sanitizable Signature Scheme. International Conference on Information Security and Cryptology (ICICS 2005), LNCS 3783, 2005, p.72-83.
5) T. Yoshioka and M. Takenaka: The realization of Partial Original Information Assurance technology for digital image data. Technology research report of the Institute of Electronics Information and Communication Engineers, ISEC2005-69, 2005, p.183-188.

Toshihiro Enami, *Fujitsu Research Institute.*
Mr. Enami received the B.A. degree in Archaeology from the University of Tokyo, Tokyo, Japan in 1981. He joined Fujitsu Limited in 1981, where he was engaged in development and implementation of IT systems and technologies for local governments. In 1996, he moved to Fujitsu Research Institute, where he has since been engaged in promotion of electronic government, reengineering of local governments, and development of IT plans for local governments.

FUJITSU Sci. Tech. J., **43**,2,(April 2007)

167