Anonymization Infrastructure for Secondary Use of Data

Yuichi Nakamura, Kanae Matsui, Hiroaki Nishi
Graduate School of Science and Technology, Keio University, Japan
{nakamura, matsui}@west.sd.keio.ac.jp, west@sd.keio.ac.jp

***Abstract*— Data containing sensitive and personal information is critical to the functioning of institutions in numerous fields, such as medical, transportation, and government. Moreover, these types of data are gaining value for secondary uses, such as market research, estimation of a route of infection, and traffic pattern analyses. From a privacy preservation viewpoint, publishing the raw data may raise significant issues because of the sensitive nature of the relevant data. Therefore, an infrastructure for publishing sensitive data while protecting privacy is required, to enable secondary use of the data. In this paper, we propose an infrastructure that supports secondary use of sensitive data in a secure manner. The proposed infrastructure preserves privacy by utilizing anonymization to publish the data; furthermore, the anonymizing process employs both a publishing rule and a request rule, thereby enhancing security. Additionally, a format for publishing datasets and their privacy-preserving rules is proposed, and is termed the XML-based Anonymize Sheets (XAS). The publishing organization and the secondary consumer of the data can designate publishing permissions and requests by utilizing XAS. The proposed infrastructure prevents additional leaks of sensitive information by utilizing the previously anonymized data as a publishing history.**

Keywords*— *secondary use of data, anonymization infrastructure, XAS

I. INTRODUCTION

Various institutions such as medical facilities, transportation facilities, and government agencies must manage large amounts of data, which may include customer information, medical records, and transaction information. This data, commonly stored in electronic form, often contains sensitive personal information. These types of data are useful, and frequently necessary, to facilitate the provision of advanced services. However, stored data may contain a considerable amount of information about individuals. This may include basic information such as age, address as well as more sensitive items such as financial data, medical records, personal preferences and history of behavior. The data contain sensitive information that organizations must protect from unauthorized use.

Recently, a movement known as Linked Open Data (LOD) has attempted to facilitate sharing of these type of data, with a goal of increasing its value. As an outgrowth of LOD, another movement known as Open Government encourages citizens to monitor government activities by publishing certain government information. For example, the United States government publishes information including economic conditions and citizens' activities on Data.gov [1].

Secondary uses of data, including location information recorded by mobile phones and data from electricity smart meters, are under consideration in Japan. The location data of mobile phones will reveal the daily travels of their users. For example, some car navigation systems utilize mobile phones to connect to datacenters, and therefore, it can obtain the car's location and other relevant data. The primary purposes of these data are to track the requirements of car's maintenance and to facilitate road services for drivers. By analyzing the data, it is possible to obtain the driving speed and location of the car. In addition, analysis of this data can identify intersections where drivers frequently brake in a sudden manner. Utilizing this information, a road maintenance squad can check the intersection, where they may identify problems such as hidden or missing signs. Data from a smart meter can provide information about the daily activities of a household. Remote observation services that monitor elderly parents attract significant attention in an aging society. These examples demonstrate that the secondary use of data can potentially create new services while enhancing the data's value. From numerous viewpoints, the secondary use of data is under consideration, and its demand is increasing.

In equal measure, this secondary use of data can result in privacy problems. In the previous examples, the location data produced by a smartphone reveals the user's location at a given time. The amount of electricity usage recorded by smart meters may reveal excessive power consumption by a household, potentially revealing their high-income status. Moreover, it is simple to publish sensitive data utilizing the Internet without proper regard to privacy. If access to this information is not adequately restricted, it may promptly result in its unauthorized use. Aside from its usefulness, publishing the data may result in the infringement of privacy rights. Therefore, techniques for publishing the data while simultaneously protecting privacy are required for the safe secondary use of the data.

To address this problem, Privacy-Preserving Data Mining (PPDM) [2, 3] and Privacy-Preserving Data Publishing (PPDP) [4, 5] are proposed. These techniques have the ability to mine or publish the data without personally identifiable information, thereby protecting privacy. Anonymization is a practical technology that supports privacy protection[5]. Anonymization technology can adjust to different privacy protection levels, thus providing flexible privacy protection. A considerable variety of studies on this technique have been performed owing

TABLE 1 MEDICAL RECORD ($\boldsymbol{k} = \mathbf{2}$)

	Birth	Gender	ID	Problem
$t1$	1970	male	121	cold
$t2$	1970	male	121	obesity
$t3$	1970	male	121	diabetes
$t4$	1980	female	121	diabetes
$t5$	1980	female	121	obesity
$t6$	1981	male	125	diabetes
$t7$	1981	male	125	cold

TABLE 2 ANONYMIZED MEDICAL RECORD ($\boldsymbol{k} = \mathbf{3}$)

	Birth	Gender	ID	Problem
$t1$	1970	male	121	cold
$t2$	1970	male	121	obesity
$t3$	1970	male	121	diabetes
$t4$	198*	human	12*	diabetes
$t5$	198*	human	12*	obesity
$t6$	198*	human	12*	diabetes
$t7$	198*	human	12*	cold

TABLE 3 ANONYMIZED MEDICAL RECORD ($\boldsymbol{k} = \mathbf{3}$)

	Birth	Gender	ID	Problem	
$t1$	1970	female	121	cold	Alice
$t2$	1970	female	121	cold	
$t3$	1970	female	121	cold	
$t4$	198*	human	12*	poor circulation	
$t5$	198*	human	12*	poor circulation	
t6	198*	human	12*	headache	Bob
t7	198*	human	12*	headache	

to its high versatility. It is one of the most preeminent privacy protection technologies in current use. Generalization and deletion of the data are necessary to prevent privacy infringements. However, they reduce the value of the data. As a result, there is a trade-off relationship between privacy protection and the utilization of the data.

Although techniques such as PPDM and PPDP have been investigated in numerous studies, a method of securely publishing the data to enable secondary use has not been definitively established. Furthermore, after calculating and publishing anonymized data from a data source, another anonymized data set, calculated and published from the same source may cause a privacy information leak if an unauthorized person can access both sets of anonymized data. When calculating and publishing anonymized data, it is necessary to consider all of the previously published data from the same source.

Considering these issues, it is crucial to establish a clear suggestion of technological guidance, an infrastructure, and a technical standard of protocols for the secondary use of data. The development of the protocol and infrastructure is especially important to its development. It will facilitate collaboration between organizations that produce the data and the companies that require the data for secondary use, and thus increase their data publishing activity. It will develop the market for secondary uses of data in conjunction with advanced services such as market research, estimation of a route of infection, and traffic pattern analysis. Moreover, it will reduce the utilization costs for both providers and consumers of secondary use data, owing to the unification of data processing procedures.

In this study, we propose a data-publishing infrastructure for secondary data use in conjunction with privacy protection by utilizing anonymization. In addition, we propose a protocol and XML-based data format for the proposed infrastructure. The infrastructure prevents further leaks of private information by employing the previously anonymized data as a publishing history.

This paper is arranged as follows. Features of anonymization and the associated privacy protection levels are described in Section 2. The design of the data format, protocol, and infrastructure for secondary data use is proposed in Section 3. The implemented mechanism is explained in Section 4. The evaluations of the proposed infrastructure are described in Section 5. Finally, we conclude the paper in Section 6.

II. ANONYMIZATION

Anonymization is one of the methods included in PPDM and PPDP. This method protects sensitive information by masking or generalizing the sensitive data. In addition,, it allows the adjustment of the privacy protection level. There are several generalization methods available for anonymization. In the following paragraphs, two relatively basic and frequently referenced generalization methods, k -anonymity and l -diversity are explained.

A. k-anonymity

K -anonymity is one of the methods utilized for generalization,[6] and it is the base of l-diversity. Further explanation of this method will incorporate the various definitions listed below.

(i) Data table

In this paper, a data list similar to a database table is termed a “data table.” Its column is termed an “attribute.” Address, birth, and gender are examples of attributes. One group of data corresponding to person or group of people is termed a “data set” and one data set is termed a “tuple.”

(ii) Attribute

An attribute among a group of related attributes that can identify a corresponding person by itself, such as name or unique ID, is termed an “identifier,” and others that cannot identify a group on their own, however, it can provide identification when combined with other attributes, such as illness, birth, gender, is termed a “quasi-identifier.”

(iii) Sensitive attribute:

A significant attribute for secondary use is termed a “sensitive attribute,” which can be selected from attributes that are not identifiers. The method will exclude this attribute from masking or generalization by anonymization. Furthermore,

tuple groups that have the same quasi-identifier values are termed "q*-block."

The definition of k-anonymity is as follows: "*In each q*-block in the data table, at least k tuples are included.*"

TABLE 1 represents an example of a medical records data table. In this table, the sensitive attribute is "Problem" and the quasi-identifiers are "Birth," "Gender," and "ID." The data consists of a $t1$~$t3$ q*-block, a $t4$, $t5$ q*-block, and a $t6$, $t7$ q*-block. It represents $k = 2$. Even if an attacker attempts to ascertain a specific individual's problem and has already obtained the individual's quasi-identifier, the attacker can narrow the results down to only two tuples. **TABLE 2** indicates that the anonymization results from **TABLE 1** are $k = 3$. The results displayed in this table demonstrate that anonymization methods provide the required privacy protection level, utilizing masking or generalization.

TABLE 4 MEDICAL RECORD ($\boldsymbol{k = 1}$)

	Birth	Gender	Problem
$t1$	1970	male	cold
$t2$	1970	male	obesity
$t3$	1970	male	diabetes
$t4$	1981	male	diabetes
$t5$	1981	female	obesity
$t6$	1982	female	diabetes
$t7$	1982	female	cold

TABLE 5 ANONYMIZED MEDICAL RECORD ($\boldsymbol{k = 2}$)

	Birth	Gender	Problem
$t1$	1970	male	cold
$t2$	1970	male	obesity
$t3$	1970	male	diabetes
$t4$	1981	human	diabetes
$t5$	1981	human	obesity
$t6$	1982	female	diabetes
$t7$	1982	female	cold

TABLE 6 ANONYMIZED MEDICAL RECORD (1) ($\boldsymbol{k = 3}$)

	Birth	Gender	Problem
$t1$	19*	male	cold
$t2$	19*	male	obesity
$t3$	19*	male	diabetes
$t4$	19*	male	diabetes
$t5$	198*	female	obesity
$t6$	198*	female	diabetes
$t7$	198*	female	cold

TABLE 7 ANONYMIZED MEDICAL RECORD (2) ($\boldsymbol{k = 3}$)

	Birth	Gender	Problem
$t1$	1970	male	cold
$t2$	1970	male	obesity
$t3$	1970	male	diabetes
$t4$	198*	human	diabetes
$t5$	198*	human	obesity
$t6$	198*	human	diabetes
$t7$	198*	human	cold

As displayed in these tables, the masking or generalization processes prevent an attacker from identifying a specific person. There are several algorithms for calculating masking or generalization. The most popular algorithm is the heuristic searching method, utilizing double-nested loops.

B. l-diversity

l-diversity is a method designed to protect the privacy of data [7]. This method considers the diversity of sensitive attributes, and it is therefore different from k-anonymity.

The definition of l-diversity is as follows: "*In all q*-blocks in a data table, there are at least l different sensitive attributes.*"

Researchers designed this method to provide protection from the following attacks.

(i) Homogeneity attack:

TABLE 3 is additional example of a medical record data table. In this case, if an attacker has acquired Alice's quasi-identifier, the attacker can read Alice's problem from this table, because no diversity exists for the sensitive attributes in the q*-block.

(ii) Background knowledge attack:

Although the $t4$~$t7$ q*-block in **TABLE 3** has a diversity of sensitive attributes, if the probability of poor circulation is very low for males and an attacker is aware of that, the attacker can read Bob's problem from **TABLE 3**.

l-diversity provides more security than k-anonymity for preserving privacy. However, the calculation cost of l-diversity is higher than k-anonymity.

III. SECONDARY USE INFRASTRUCTURE

The demand for the secondary use of the data such as medical records is increasing, because it may enable the estimation of infection routes. However, medical data frequently includes sensitive and private information. The medical data providers should define the anonymization methods and the related privacy protection levels when publishing the data. In addition, when the data provider permits several methods of anonymization, the consumers of the data must select a method that matches their requirements. Moreover, consumers of the anonymized data should avoid obtaining private data that exceeds their requirements, including situations where the data provider permits the lower protection level and thus provides the private data. Therefore, the anonymization data infrastructure should provide a method to define anonymization methods and protection levels that fulfill the requirements for both data providers and data consumers.

To meet these requirements, data publishing with anonymization is required. However, PPDP utilizing anonymization has numerous problems. One of the problems is

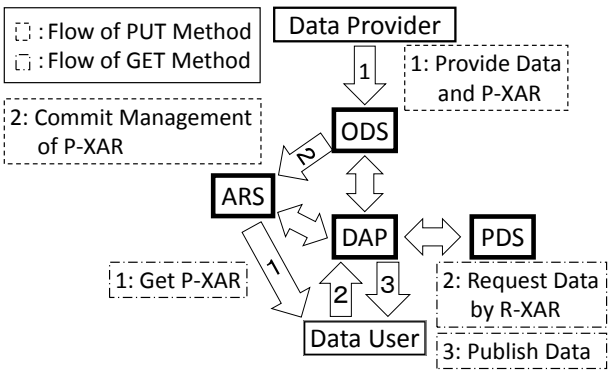

Fig. 1 Overview of proposed infrastructure

that no protocols and formats currently exist to enable secure data publishing, as described in the introduction. The other is loss of anonymity by publishing the same data multiple times. **TABLE 4** is an example of a medical record data table. **TABLE 5** is an anonymized $k = 2$ data table with data from **TABLE 4**, and **TABLE 6** is another anonymized $k = 3$ data table with data from **TABLE 4**. In this case, those who can obtain both the anonymized data of $k = 2$ and $k = 3$ can obtain the $k = 1$ data, including situations where the data provider did not permit the publishing of $k = 1$ data. This results in leak of privacy information. One cause of this problem is that previously published data is not referenced in the anonymization process;

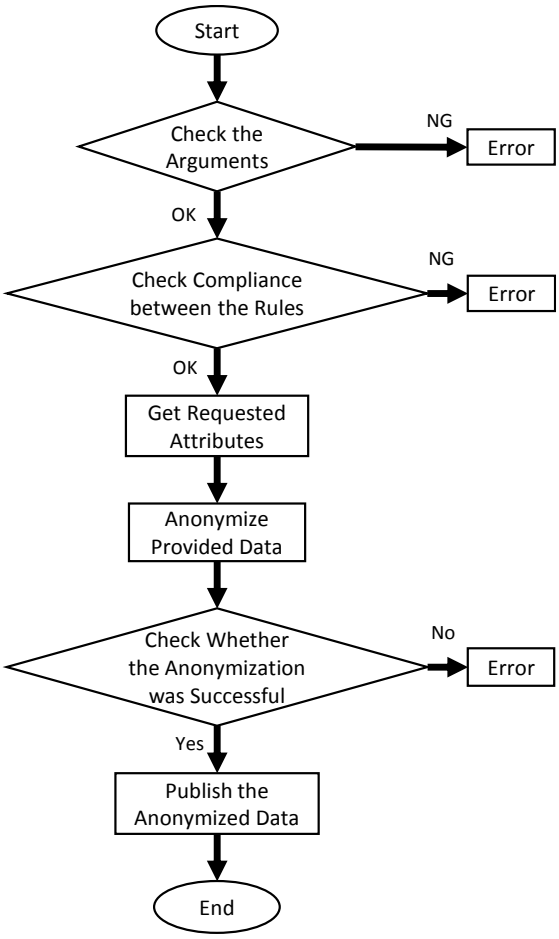

Fig. 2 Flowchart of the Implemented Mechanism

as a result the coherence between the $k = 2$ and $k = 3$ data was severed. **TABLE 7** is another example of a $k = 3$ data table. Utilizing **TABLE 7** instead of **TABLE 6** avoids the problem described above. **TABLE 7** was generated by anonymizing **TABLE 5** instead of anonymizing **TABLE 4**, to maintain coherency in masking and generalization. The proposed anonymizing process can prevent further leaks of privacy information.

To address these problems, we proposed a data-publishing infrastructure. It manages the previously published data for the anonymization without the loss of anonymity, and provides safe secondary use and anonymization. For encryption technology, it utilizes Public Key Infrastructure (PKI). Certificate Authority serves a function as an authorized organization for certifying the public key of servers on the Internet. For this discussion, the anonymization technology and proposed infrastructure can be associated with the encryption technology and PKI, respectively.

A. Design of secondary use infrastructure

The proposed infrastructure can be divided into four organizations as follows.

(i) Original data storeroom organization (ODS)

This organization manages data provided by the data folder. The data folder is considered the data provider when the data is managed by ODS. When providing data to ODS, the data folder prepares data for publishing and provides an allowance rule by utilizing a specially designed format. This format is termed XML-based Anonymization Sheets (XAS). The details of XAS are described in the following section. Publishing rule descriptions utilize a subset of XAS, termed XML-based Anonymization Rules (XAR). The data folder generates data as D-XAS and the publishing rules (P-XAR) correspond to the D-XAS. D-XAS should include the link to the P-XAR. ODS should be responsible for maintaining the original data written as D-XAS in a secure manner. This data registration process is based on the PUT method.

(ii) Anonymizing rules storeroom organization (ARS)

This organization manages P-XAR. P-XAR will be openly published for users who need to access anonymized data based on the original data. P-XARs stored in the ARS can exhibit data when it is available for its secondary use. A P-XAR is stored by utilizing a PUT method issued by ODS.

(iii) Data anonymizing and publishing organization (DAP)

This organization anonymizes the original data (D-XAS) based on a publishing rule (P-XAR) and a request rule (R-XAR). A secondary use data consumer generates an R-XAR and provides it to the DAP. An R-XAR contains relevant information for D-XASs such as a URL, the requested anonymization method, its privacy level and anonymization range required to obtain the data for secondary use. The DAP receives the header of the requested D-XAS to access the link of the R-XAR. This header information does not include data. This header information is also described by using an XAR termed H-XAR; the DAP verifies its compliance by checking with the R-XAR and P-XAR requested from the ARS, according to the H-XAR. In this process, a user utilizes a GET

method in conjunction with the R-XAR option. If it returns a compliance error, the user receives an appropriate error message. This message utilizes the HTTP error message protocol. If no error occurs, DAP issues a GET message to obtain the D-XAS from the ODS, and issues a subsequent GET message to receive the published XAS (P-XAS) from the PDS. The PDS is described in the following paragraph (iv). The DAP generates P-XASs as anonymized data, and the response from the R-XAR of the user. The user receives the anonymized data resulting from the GET method. Finally, the DAP stores the generated P-XAS issues by utilizing the PUSH method. This P-XAS is utilized to prevent further privacy leaks.

(iv) Published data storeroom organization (PDS)

This organization manages data previously published by the DAP as P-XASs. It may store all anonymized data generated by the DAP. However, to optimize data storage capacity, it is sufficient for the PDS to store only one P-XAS as anonymized data for each D-XAS, according to the one-direction anonymization policy. When generating P-XASs from D-XASs according to the requested R-XAR, it is sufficient to generate P-XASs according to the R-XAR, and store the P-XAS to the PDS. However, when generating another P-XAS from the same D-XAS according to another R-XAR, the DAP should obtain all P-XASs related to the D-XAS from the PDS. The DAP should consider all of these P-XASs when generating new P-XASs to observe P-XARs. Therefore, we propose one-directional anonymization to avoid this process. The process is as follows.

```
 1 <?xml version="1.0" encoding="utf-8"?>
 2 <?xml-anonymize type="text/xas" href="p-xar.xas"?>
 3 <list>
 4  <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:v="http://www.w3.org/2006/vcard/ns#">
 5   <v:Kind rdf:about = "http://foo.com/me/hogehoge" >
 6    <v:fn>Hoge Foo</v:fn>
 7    <v:bday>1980-01-01</v:bday>
 8    <v:hasTelephone>
 9     <rdf:Description>
10      <rdf:value>+81-45-566-1454</rdf:value>
11      <rdf:type
rdf:resource="http://www.w3.org/2006/vcard/ns#Work"/>
12      <rdf:type
rdf:resource="http://www.w3.org/2006/vcard/ns#Voice"/>
13     </rdf:Description>
14    </v:hasTelephone>
15    <v:hasAddress>
16     <rdf:Description>
17      <v:street-address>123-45 Hoge Village</v:street-address>
18      <v:locality>FooCity</v:locality>
19      <v:postal-code>5555</v:postal-code>
20      <v:country-name>Japan</v:country-name>
21     </rdf:Description>
22    </v:hasAddress>
23   </v:Kind>
24  </rdf:RDF>
25  <OfficeScale>100ha</OfficeScale>
26  <PowerConsumption>10kWh</PowerConsumption>
27
28  <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:v="http://www.w3.org/2006/vcard/ns#">
29   <v:Kind rdf:about = "http://foo.com/me/db" >
```

Fig. 3 D-XAS Example (Extract)

(i) The DAP generates P-XASs according to P-XARs, instead of R-XARs, and stores it in the PDS. Therefore, the PDS stores the anonymized data, and it is anonymized according to the declared level in P-XAR. This P-XAS is not sent to the users if the requested level in the R-XAR is higher than the level in the P-XAR; this indicates the $k$value is larger than that of the P-XAR in k-anonymity.

(ii) DAP generates P-XASs according to the R-XARs. In this generation, the DAP only uses the first P-XAS generated from the P-XAR. DAP generalizes new P-XASs by adding "wild cards" as masking from the initial P-XAS. The DAP does not remove any of the "wild cards" provided as masking in the first P-XAS. Therefore, a one-directional anonymizing process should be considered.

(iii) The DAP can generate any type of P-XAS that satisfies both the R-XAR and the P-XAR by following the process described in (i) and (ii). In a scenario where k-anonymity and l-diversity are mixed, it is sufficient to generate a P-XAS that has a lower anonymization level than k-anonymity and l-diversity. For example, assume that 3-anonymity and 3-diversity are permitted in P-XARs, and 4-diversity is requested by R-XAR. In this case, DAP generates the initial P-XAR by utilizing 3-anonymity. The DAP can generate any type of P-XAR by utilizing the initial P-XAR, according to the one-directional anonymizing process.

To enable the data transfer between these organizations, data providers and data consumers will utilize SSL and PKI if they transfer the data over the Internet. In the following discussions, four organizations are exhibited in order to clarify each role. It is possible to merge some of them into a single organization.

Fig. 1 represents proposed organizational structure and data connections between the organizations.

IV. XML-BASED ANONYMIZE SHEETS (XAS)

We propose XML-based Anonymization Sheets (XAS) as a format to define the rules and data descriptions. To distinguish the rules from the data, XML-based Anonymization Rules (XAR) are also proposed as a subset of XAS. XAS and XAR differ because XAR does not contain data as contents. All transactions in the proposed infrastructure utilize the XAS and its subset, XAR. XAS is designed according to Extensible Markup Language (XML). **Fig. 3** lists an example of D-XAS. It includes the information to enable anonymization, including combinations of the sensitive attribute names and quasi-identifiers, permitted anonymization methods and levels, and data attributes such as created date, updated date and history, ownership, copyrights, comments, and others. **Fig. 4** lists an example of a P-XAR. It does not contain raw data; it only declares the required anonymization methods and levels. To enable masking or generalization processes, it can define the delimiter for distinguishing data sections. In this example, "BirthDay" is split utilizing the '-' character. During the anonymizing process, the character is used to define the generalization boundary. If the data employs a general and standardized format, for example, BirthDay should be separated by '-,' it can generalize the data entry by referring to the default rule. As an additional feature, the data provider may

```
<?xml version="1.0" encoding="utf-8"?>
<anonymize>
 <head>
  <publishacceptance sensitive="divisional" quasi="divisional" />
  <firstdatasetposition>
   <list>
    <rdf:RDF />
   </list>
  </firstdatasetposition>
  <sensitive type="k(>=3), l(>=2)">
   <rdf:RDF>
    <v:Kind>
     <v:hasTelephone>
      <rdf:Description>
       <rdf:type number="2" />
      </rdf:Description>
     </v:hasTelephone>
    </v:Kind>
   </rdf:RDF>
   <PowerConsumption />
  </sensitive>
  <sensitive type="k(>=3), l(>=2)">
   <OfficeScale />
  </sensitive>
  <group name="addr" type="quasi" level="k(>=3), l(>=3)"/>
 </head>
 <rdf:RDF>
  <v:Kind>
   <v:fn note="Full Name" />
   <v:bday note="BirthDay" type="quasi" level="k(>=2)" sprit="-" />
   <v:hasTelephone>
    <rdf:Description>
     <rdf:value note="TelephoneNumber" type="open" sprit="\s" />
     <rdf:type note="Number Type" attribute="rdf:resource"
number="2" />
    </rdf:Description>
   </v:hasTelephone>
   <v:hasAddress>
    <rdf:Description note="Addresses">
     <v:street-address group="addr" priority="4" />
     <v:locality group="addr" priority="3" />
     <v:postal-code group="addr" priority="2" />
     <v:country-name group="addr" priority="1" />
    </rdf:Description>
   </v:hasAddress>
  </v:Kind>
 </rdf:RDF>
 <OfficeScale note="OfficeScale" />
 <PowerConsumption type="open" note="PowerConsumption" />
</anonymize>
```

Fig. 4 P-XAR Example

publish data samples without data publishing limits to publicize the data's availability. This open information is termed "open attribute." This open attribute can be declared in a data entry.

The secondary data user can request access to the open attributes by utilizing R-XAR. **Fig. 5** lists an example of an R-XAR. If the secondary data consumer requests attributes identified as quasi-identifiers, DAP publishes anonymized data that contains attributes calculated as quasi-identifiers. The user also declares the required anonymization method, privacy protection level, sensitive attributes combinations, open attributes, and quasi-identifiers utilizing the R-XAR.

The formats of XAS and its subset XAR utilize the Cascading Style Sheets (CSS) format and the Semantic Web standard. The XAS can be processed utilizing an XML schema, RDL schema, OWL method, and other related tools.

V. Implementation of PDS

We implemented a DAP application to verify its performance and feasibility. The DAP is the most complicated application, and must be implemented first to enable the evaluation of the proposed infrastructure. The implemented application can confirm the compliance of P-XARs and R-XARs as publishing and requesting rules, respectively. The DAP application utilizes TinyXML-2 [8] for parsing the XAS. P-XARs for the anonymized data were stored in PostgreSQL, a prominent database management system. The application can anonymize data according to the k-anonymity and l-diversity anonymization methods.

Fig. 2 displays a flowchart for the implemented application. Initially, it receives original data as a D-XAS, its publishing rules as a P-XAR, and its requesting rules as an R-XAR. It verifies the format and compliance between the P-XAR and the R-XAR. Subsequently, the application receives the requested data according to the published P-XAR from PostgreSQL, if the data exists. The program then initializes the process of anonymization described in Section II. This program also emulates the process of one-directional anonymization described in Section III. Finally, the program stores the anonymized data as P-XAS into the database.

VI. Evaluation

To evaluate the DAP application, we assumed a typical application example as follows. We utilized the Web access history captured by the designed packet-capturing software implemented on our lab's gateway server. It captures all

```
<?xml version="1.0" encoding="utf-8"?>
<anonymize type="k(3)">
 <head>
  <sensitive>
   <rdf:RDF>
    <v:Kind>
     <v:hasTelephone>
      <rdf:Description>
       <rdf:type number="2" />
      </rdf:Description>
     </v:hasTelephone>
    </v:Kind>
   </rdf:RDF>
   <PowerConsumption />
  </sensitive>
  <group name="addr" type="quasi" />
 </head>
 <rdf:RDF>
  <v:Kind>
   <v:bday />
   <v:hasTelephone>
    <rdf:Description>
     <rdf:value note="TelephoneNumber" type="quasi" />
    </rdf:Description>
   </v:hasTelephone>
  </v:Kind>
 </rdf:RDF>
 <PowerConsumption note="PowerConsumption" />
</anonymize>
```

Fig. 5 R-XAR Example

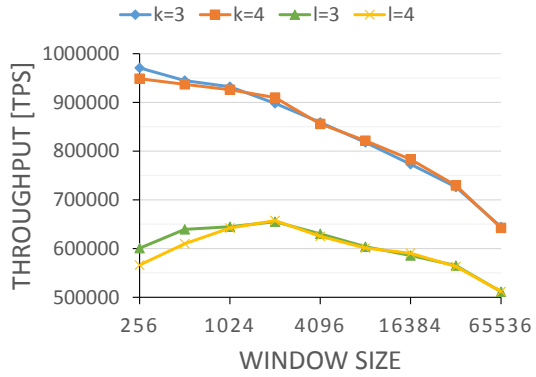

Fig. 6 Throughput of the Implemented Program (1)

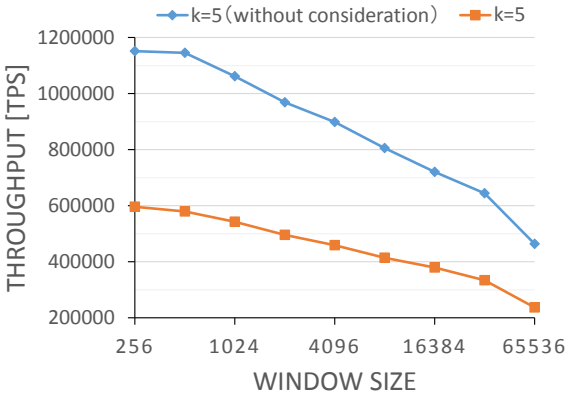

Fig. 7 Throughput of Implement Program (2)

transactions between the Internet and our intranet. Data representing the relationship between the users' IP addresses and the accessed URLs was captured, by referencing the access history. Utilizing this information, we can track the users who frequently access certain URLs. This sample application should manage this captured data as anonymized information. Utilizing this information, we can provide new services, such as a trend survey of Web accesses, a Web page recommendation service, and other useful services.

The application was implemented in C++, utilizing the Debian 7.0 operating system. The host PC includes an Intel Xeon X5650 2.67 GHz CPU, and 36 GB of DDR3 Memory. Test data included two months of Web page access history. The application utilized the first 65,536 instances of all captured accesses, comprising 6,294,273 bytes. In this data, the sensitive attribute is the domain name of Web sites, and quasi-identifier is the user's IP address.

Fig. 6 displays the relationship between the window size and the anonymization throughput of the implemented application. In this example, one-direction anonymization was not utilized; this increases the application's potential vulnerability to privacy breaches. However, the calculation cost is lower, because it calculates k-anonymity or l-diversity only once during the initial anonymizing process. This evaluation measured throughput in tuples per second (TPS). Applying a window size of 4,000, the program can anonymize approximately 600,000 tuples to 800,000 tuples per second. This demonstrates the application is capable of processing large amounts of data.

Fig. 7 additionally displays the relationship between window size and throughput. This diagram, displays the throughput for one-direction anonymization, along with throughput results where one-direction anonymization is not considered. Loss of anonymity was not observed in cases where one-direction anonymization was utilized. However, anonymization throughput was reduced by half when one-direction anonymization was considered; when one-direction anonymization is considered, the program performs the anonymization calculation twice.

VII. Conclusion

We proposed an infrastructure to facilitate the secondary use of data. The proposed infrastructure manages anonymization and the relevant data. We displayed the organizational structure of the proposed anonymization infrastructure and data transactions. This infrastructure prevents further leaks of private information by utilizing one-directional anonymization. We also proposed XAS and its subset, XAR as a format. A protocol to exchange XAS and XAR was also described. The proposed infrastructure will facilitate future services related to the secondary use of data. To evaluate our proposed solution, we implemented an application for DAP. This evaluation demonstrated that the proposed application can anonymize from 600,000 to 800,000 tuples per second, and it exhibits sufficient performance to process large amounts of data.

Acknowledgement

This work was partially supported by Funds for integrated promotion of social system reform and research and development, MEXT, Japan, and by MEXT/JSPS KAKENHI Grant (B) Number 24360230 and 25280033.

References

[1] U.S. Federal Government. DATA.GOV. http://www.data.gov/ (2014.5).

[2] Rakesh Agrawal; Ramakrishnan Srikant, "Privacy-preserving data mining," SIG-MOD, Vol. 29, pp. 439-450, 2000.

[3] Yehuda Lindell; Benny Pinkas, "Privacy Preserving Data Mining," Journal of Cryptology, Vol. 15, pp. 177-206, 2002.

[4] Bee-Chung Chen; Daniel Kifer; Kristen LeFevre; Ashwin Machanavajjhala, "Privacy-Preserving Data Publishing," Foundations and Trends in Databases, Vol. 2, No. 1-2, pp. 1-167, 2009.

[5] Benjamin C. M. Fung; Ke Wang; Rui Chen; Philip S. Yu, "Privacy-preserving data publishing: A survey of recent developments," ACM Computing Surveys (CSUR), Vol. 42, No. 4, 2010.

[6] Latanya Seeney, "k-anonymity: a model for protecting privacy," International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, Vol. 10, No. 5, pp. 557-570, 2002.

[7] Ashwin Machanavajjhala; Daniel Kifer; Johannes Gehrke; Muthuramakrishnan Venkitasubramaniam, "L-diversity: Privacy beyond k-anonymity," ACM Transactions on Knowledge Discovery from Data (TKDD), Vol. 1, No. 1, 2007.

[8] TinyXML-2, http://www.grinninglizard.com/tinyxml2/ (2014.5).