Data-Purpose Algebra: Modeling Data Usage Policies

Chris Hanson (cph@csail.mit.edu) Tim Berners-Lee (timbl@w3.org)
Lalana Kagal (lkagal@csail.mit.edu) Gerald Jay Sussman (gjs@mit.edu)
Daniel Weitzner (djweitzner@csail.mit.edu)

Massachusetts Institute of Technology
Computer Science and Artificial Intelligence Laboratory
Cambridge, MA, USA

Abstract

Data is often encumbered by restrictions on the ways in which it may be used. These restrictions on usage may be determined by statute, by contract, by custom, or by common decency, and they are used to control collection of data, diffusion of data, and the inferences that can be made over the data. In this paper, we present a data-purpose algebra that can be used to model these kinds of restrictions in various different domains. We demonstrate the utility of our approach by modeling part of the Privacy Act (5 USC §552a)[1], which states that data collected about US citizens can be used only for the purposes for which it was collected. We show (i) how this part of the Privacy act can be represented as a set of restrictions on data usage, (ii) how the authorized purposes of data flowing through different government agencies can be calculated, and (iii) how these purposes can be used to determine whether the Privacy Act is being enforced appropriately.

1. Introduction

Privacy protection in large scale, decentralized information systems such as the Web and large commercial data mining applications requires new technical tools and public policy approaches [15]. In particular, we believe that new systems approaches are needed to enable assessment of accountability to support policies that govern *uses* of personal information. To this end, we describe here a data-purpose algebra that enables the expression of data-purpose and usage restrictions. Our goal is to build systems in which the specific uses of personal data are transparent to authorized observers and are subject to effective accountability assessments by those who seek to assure privacy policy compliance.

Data about individuals can be obtained easily through a variety of ways including tracking user behavior on websites, monitoring purchases on credit cards, querying government and commercial databases, and even using environmental sensors. This data can be used to make inferences (of uncertain accuracy) about the individuals for a range of purposes from market research and customized marketing to verifying the time during which the individual was at work. These inferences can also lead to adverse consequences. For example, a person might be incorrectly identified as a terrorist and prevented from boarding a plane. However, there are usually restrictions on the ways data can be collected and used. These encumbrances may be determined by statute, by contract, by custom, or by common decency. Some of these restrictions are intended to control the diffusion of the data, while others are intended to delimit the consequences of actions predicated on that data.

Data can also be sent from one individual or organization to another. In this case, the allowable uses of data may be further restricted by the sender: "I am telling you this information in confidence. You may not use it to compete with me, and you may not give it to any of my competitors." Data may also be restricted by the receiver: "I don't want to know anything about this that I may not tell my spouse."

Although the details may become complex as data is passed from one individual or organization to another, the restrictions on the uses to which the data may be put are changed in ways that can often be formulated as algebraic expressions. These expressions describe how the restrictions on the use of a particular data item may be computed from the history of its transmission: the encumbrances that are added or deleted at each step. A formalization of this process is a *Data-Purpose Algebra* description of the process.

One pervasive assumption behind our formalization is that the restrictions on a data item do not depend solely on the content of the item, but rather on the content and

[1] For an overview of the Privacy Act of 1974 see the website: http://www.usdoj.gov/oip/04_7_1.html

its annotated provenance. For example, a law-enforcement official may not act on improperly obtained evidence, but if the same information were redundantly obtained through lawful channels the official may act, assuming that the independence meets certain constitutional requirements.

2. Data-Purpose Algebra

To formally describe the ways that the use of data may be restricted and the way in which the restrictions are transformed as the data is processed and passed from one agent to another, we annotate each data item with extra information. Each data item i, in addition to its content $q = Q_{\text{D}}(i)$, is annotated with its agent $a = A_{\text{D}}(i)$, a category $k = K_{\text{D}}(i)$, and a set of purposes $p = P_{\text{D}}(i)$ for which it can be used. An item is constructed from its content, agent, category, and purposes $i = I(q, a, k, p)$. The agent is the producer of this data. The category is a set of data items containing this particular item. This set may be named but is not likely to be enumerated; for example a typical legal category is "US citizen." The set of purposes is explicit; a typical purpose in the set of purposes is "criminal law enforcement."

A data item i may be processed by some agent a' to produce a new data item. (See figure 1.) The new data item has the same kinds of annotations as the original one, but the process generates new content and new annotations as functions of the original. The functions are specific to the kind of process performed by the agent.

For example, medical data about an identified person may be severely restricted as to its allowable uses. But it may be anonymized for use in the education of medical doctors. In such a case, the allowed purposes of the anonymized data may be wider than those of the original data, and the category of the data will be different. On the other hand, when a person enters a medical establishment for treatment a record is made of the patient's name, address, and date of birth. This data itself typically has few usage restrictions, but the fact of it appearing in a medical record adds restrictions required by the HIPAA law.

An agent may combine data from multiple agents to produce new data. (See figure 2.) In this case the functions may be considerably more complex. For example, a person at the medical office may use a public source, such as a telephone directory, to verify the recorded telephone number of a patient. This process combines highly restricted information from a medical record with unrestricted public information, but the result remains restricted.

3. An Example Formalization: Privacy Act

In the illustration that follows we consider a simplified formulation of the rules for data passed among Systems of Records (SORs), specified by the associated Systems of Records Notices (SORNs), as defined by the Privacy Act (5 USC §552a).

Let r be a data repository; if the repository is a SOR, it has an associated SORN $n = N(r)$, which gives information about the permissible uses of the data in the SOR. The SORN specifies input conditions: the allowed sources $O_{\text{S}}(n)$ from which data may be collected, the data categories $K_{\text{S}}(n)$ that may be collected, and the purposes $P_{\text{S}}(n)$ for which data that is collected may be used. It also specifies a set of routine uses $U(n)$ for data extracted from that SOR.

Each routine use $u \in U(n)$ specifies a set of possible recipient organizations $O_{\text{R}}(u)$, categories of data $K_{\text{R}}(u)$ that may be transferred to those organizations, and the set of authorized purposes $P_{\text{R}}(u)$ for which the specified recipient organizations may use the given data. Any particular recipient r_1 may be a sub-organization of a possible recipient organization r_2 specified in a SORN. This (non-strict) relation is notated $r_1 \preceq r_2$. Likewise, any particular category k_1 may be a subset of a category k_2 specified in a SORN. This relation is notated $k_1 \subseteq k_2$.

The purposes allowed for data i that has been transferred from a SOR s to a SOR r depend on the purposes that came with the data and the input conditions on the SORN for r. So, if s is not one of the allowed sources or the category of the data is not one of the allowed categories the data may not be used for any purpose:

$$\begin{aligned} R_{\text{IN}}(i, s, r) = P_{\text{S}}(r) \text{ if } & \exists o((o \in O_{\text{S}}(N(r))) \wedge (s \preceq o)) \\ & \wedge \exists k((k \in K_{\text{S}}(N(r))) \wedge (K_{\text{D}}(i) \subseteq k)) \\ = \{\} \text{ otherwise} & \end{aligned}$$

The set of applicable routine uses $A(i, s, r)$ for transfer of data item i from a SOR s to a recipient r is just the set of those entries for which the recipient is allowed by the SORN for s and for which the category of the data $K_{\text{D}}(i)$ is in the allowed categories $K_{\text{R}}(u)$ for that routine use u:

$$\begin{aligned} A(i, s, r) = \{u \in U(N(s)) \mid & \\ & \exists o((o \in O_{\text{R}}(u)) \wedge (r \preceq o)) \\ & \wedge \exists k((k \in K_{\text{R}}(u)) \wedge (K_{\text{D}}(i) \subseteq k))\} \end{aligned}$$

The restriction on authorized purposes of a transfer from a source to a recipient is that the purposes must be authorized by one or more of the applicable routine uses.

$$R_{\text{OUT}}(i, s, r) = \bigcup_{u \in A(i,s,r)} P_{\text{R}}(u)$$

The authorized purposes $Z(i, s, r)$ to which a recipient r may put a data item i extracted from a source s is then restricted to be those purposes particular to that data item

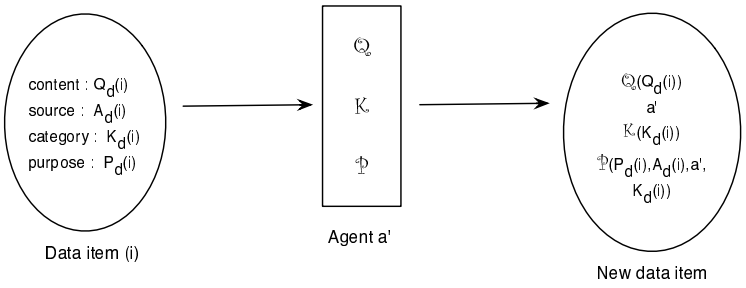

Figure 1. Unary Process: A data item i may be processed by some agent a' to produce a new data item. The new content is some function $\mathcal{Q}(Q_{\text{D}}(i))$ of the given content. The agent of the new data item is a', the new category is a function $\mathcal{K}(K_{\text{D}}(i))$ of the given category, and the allowed purposes of the new data item is a more complex function $\mathcal{P}(P_{\text{D}}(i), A_{\text{D}}(i), a', K_{\text{D}}(i))$ that may depend on the original purposes, the agents, and the category of the original data.

that are also allowed by the authorized purposes specified in the SORN:

$$Z(i, s, r) = P_{\text{D}}(i) \cap R_{\text{IN}}(i, s, r) \cap R_{\text{OUT}}(i, s, r)$$

So $Z(i, s, r)$ is the set of purposes of the new item held by the recipient r with the content of the old item i held by the source s.

The result of a transfer process A_{XFER} of an item i from a source s to a recipient r is a new item:

$$I(Q_{\text{D}}(i), A_{\text{XFER}}, K_{\text{D}}(i), Z(i, s, r))$$

4. Implementation

We are constructing a system [15] that uses the data-purpose algebra to derive usage purposes for data, and consequently analyze usage of the data by government agencies. Our system expects as input the historical log of data collection, analysis, and transfer between individuals and government agencies. Based upon current government efforts, we presume that this log as well as case activities will exist in XML [2]. We plan to automatically convert the XML transactional data into the Resource Description Framework (RDF) [9]; annotate the transactional data with agent, category, and purposes as described in this paper; and then encode the derivation steps in the Proof Markup Language [5], thereby providing an interoperable justification representation for explanation and accountability. We are also defining notices (in RDF) for the Systems of Records that appear in the transaction log. In our hypothetical scenario, a traveler named John Doe from New York boards a flight in New York and sets in motion a chain of inferences (some of which are factually incorrect and some of which are reached in violation of the Privacy Act) that generates a series of adverse consequences for him.

A transaction log that leads up to an adverse consequence—such as John Doe being arrested—is passed through the data-purpose algebra of the Privacy Act. This causes each data item in the log to be annotated with its authorized purposes, as described in the earlier section. A data item that has an empty list of purposes shows the point at which a policy violation occurred. All further uses of these data items without purposes, including inferences made, in the log are also invalid.

We have built an implementation of the data-purpose algebra in Scheme [10]. To illustrate how closely the program corresponds to the mathematics, figure 3 shows an implementation of $R_{\text{OUT}}(i, s, r)$ and $A(i, s, r)$ in Scheme. The overall correspondence between code and mathematics is quite close.

In order to enable users to understand where and how the Privacy Act has been violated in a particular transaction log, we are developing a user interface for our system. The UI provides several views into the annotated log, including: (a) a time-line of events; (b) flow of data through the records systems; (c) transaction details; and (d) data-purpose calculations.

5. Related Work

Our algebra is closely related to work on privacy policy representation and enforcement. W3C's P3P framework allows websites to publish their privacy policy, which can be matched against users' privacy preferences [16]. Complementary languages for defining privacy preferences such as APPEL [4] and XPREF [1] have also been defined. The P3P framework is specifically aimed at web privacy in terms of

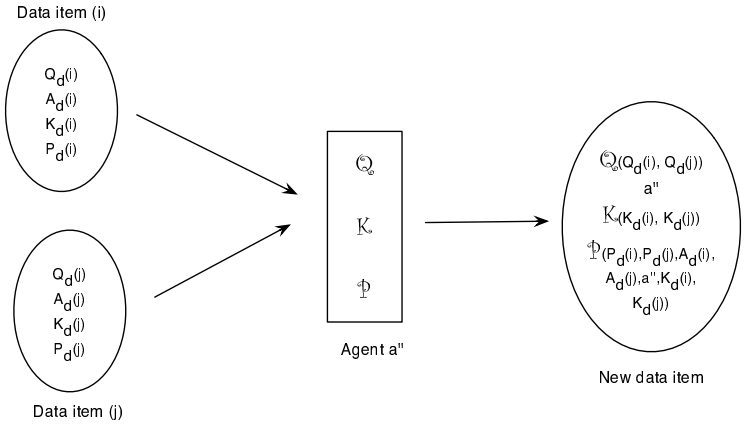

Figure 2. Binary Process: Two data items, i and j, may be processed by some agent a'' to produce a new data item. The new content is some function $\mathcal{Q}(Q_{\text{D}}(i), Q_{\text{D}}(j))$ of the content of i and j. The agent of the new data item is a'' and the new category is a function $\mathcal{K}(K_{\text{D}}(i), K_{\text{D}}(j))$ of the categories of i and j. The allowed purposes of the new data item is a complex function that may depend on the original purposes, the agents, and the categories of i and j.

collecting, storing, and sharing user information and deals solely with cookies and clickstream data. In contrast, our data-purpose algebra has a broader scope and models restrictions on how data can be used in general.

While P3P is concerned with privacy protection issues for web users, EPAL 1.1 [8] specifies enterprise-level policies for data objects in the enterprise. Like P3P, EPAL has a limited scope and does not provide a general model of data restrictions.

Policy languages such as Extensible Access Control Markup Language (XACML) [7] and KAoS [13, 14] are used to describe access control policies. Using these languages it is possible to define restrictions on actions that deal with specific data thereby enforcing restrictions on data usage. However, access control policies are enforced at the time of the user request, whereas our work is focused on accountability.

Privacy approaches such as ContextBroker [3] and Semantic eWallet [6] are being developed for pervasive computing environments where the behavior of users may be monitored and used. These approaches enforce users' privacy preferences by preventing their data from being used by or modifying the data available to context aware services.

6. Future Work

The example in this paper shows how to cover many kinds of formalizable requirements, such as those of the Privacy Act. But there are harder problems. Informal and implicit restrictions on data usage may result in contradictory conclusions about what purpose restrictions apply. Further work is required to define means by which these conflicting results can be resolved.

Anonymization is often used to remove identifying characteristics of data. However, it has been shown that sets of "anonymized" data can often be combined to discover the identities of the parties [11]. Is the combined data restricted? It depends on the laws. If the law requires that medical records are restricted, then that conclusion is independent of how they are derived. On the other hand, it is possible to combine two restricted pieces of information to produce a less restricted deduction. For example, if the *same* information is available from two different sources, then the restriction on the combination may be relaxed to be the uses allowed by each source separately, or it may not, depending on the details.

The data-purpose algebra is designed to compute allowed purposes based on how data is derived, but as in the case of re-constituted medical records, this is not always sufficient. We believe that this model can be extended to handle such cases by adding content-dependent global rules, for example by using the category of a data item to indicate when such a rule should be applied.

7. Summary

The algebraic approach is well suited to modeling the allowable uses of information when the restrictions on that

```
(define (restriction-on-output record source recipient)
  (rdf-set:union*
   (append-map ts:routine-use.purposes
               (applicable-routine-uses record source recipient))))

(define (applicable-routine-uses record source recipient)
  (keep-matching-items (ts:sorn.routine-uses (ts:sor.notice source))
    (lambda (use)
      (and (there-exists? (ts:routine-use.recipients use)
             (lambda (recipients)
               (org-elt? recipient recipients)))
           (there-exists? (ts:data-record.categories record)
             (lambda (category)
               (there-exists? (ts:routine-use.categories use)
                 (lambda (category*)
                   (rdf-subset? category category*)))))))))
```

Figure 3. Scheme implementation of $R_{\text{OUT}}(i, s, r)$ and $A(i, s, r)$.

use are determined by the path by which the information is obtained, but it is not so good at dealing with restrictions that are time dependent or inherent in the content of the information, independent of the path.

When formalized algebraically, computations are directly representable as purely functional computer programs. This makes it easy to verify that a program that implements the data-purpose algebraic computations is correct.

Acknowledgements

Hal Abelson, Deborah McGuinness, and K. Krasnow Waterman read drafts of this paper and provided many useful and insightful comments. The work reported in this paper is supported by the US National Science Foundation Cybertrust (05-518) program.

References

[1] R. Agrawal, J. Kiernan, R. Srikant, and Y. Xu. XPref: A preference language for P3P. *Comput. Networks*, 48(5):809–827, 2005.

[2] T. Bray, J. Paoli, C. M. Sperberg-McQueen, E. Maler, and F. Yergeau. Extensible markup language (XML) 1.0. http://www.w3.org/TR/REC-xml/.

[3] H. Chen, T. Finin, and A. Joshi. An intelligent broker for context-aware systems. In *Adjunct Proceedings of Ubicomp 2003, Seattle, Washington, USA, October 12–15, 2003*, Oct. 2003.

[4] L. Cranor, M. Langheinrich, and M. Marchiori. A P3P preference exchange language (APPEL). http://www.w3.org/TR/P3P-preferences/.

[5] P. P. da Silva, D. L. McGuinness, and R. Fikes. A proof markup language for semantic web services. *Information Systems*, 31(4–5):381–395, June–July 2006.

[6] F. Gandon and N. Sadeh. Semantic web technologies to reconcile privacy and context awareness. *Web Semantics Journal*, 2004.

[7] S. Godik and T. Moses. OASIS extensible access control markup language (XACML). OASIS Committee Specification cs-xacml-specification-1.0, Nov. 2002.

[8] IBM. EPAL 1.1. http://www.zurich.ibm.com/security/enterprise-privacy/epal/Specification/index.html, 2003.

[9] F. Manola and E. Miller. RDF primer. http://www.w3.org/TR/rdf-primer/.

[10] G. L. Steele, Jr. and G. J. Sussman. The revised report on scheme, a dialect of lisp. MIT AI Memo 452, Jan. 1978.

[11] L. Sweeney. Guaranteeing anonymity when sharing medical data, the datafly system. *Journal of the American Medical Informatics Association*, 1997.

[12] U. S. Department of Justice and Department of Homeland Security. National information exchange model (NIEM). http://www.niem.gov/.

[13] A. Uszok, J. Bradshaw, P. Hayes, R. Jeffers, M. Johnson, S. Kulkarni, M. Breedy, J. Lott, and L. Bunch. DAML reality check: A case study of KAoS domain and policy services. In *International Semantic Web Conference (ISWC 03), Sanibel Island, Florida*, 2003.

[14] A. Uszok, J. M. Bradshaw, R. Jeffers, M. Johnson, A. Tate, J. Dalton, and S. Aitken. Policy and contract management for semantic web services. In *AAAI Spring Symposium, First International Semantic Web Services Symposium*, 2004.

[15] D. Weitzner, H. Abelson, T. Berners-Lee, C. Hanson, J. Hendler, L. Kagal, D. McGuinness, G. Sussman, and K. K. Waterman. Transparent accountable inferencing for privacy risk management. In *The Semantic Web meets eGovernment, AAAI Spring Symposium*, 2006.

[16] World Wide Web Consortium. Platform for privacy preferences (P3P) project. http://www.w3.org/P3P/.