Data Service Agreements: Toward a Data Supply Chain

Len Seligman, Arnon Rosenthal

The MITRE Corporation
McLean, Virginia and Bedford, Massachusetts, USA
{seligman, arnie}@mitre.org

James Caverlee*

Georgia Institute of Technology
Atlanta, Georgia, USA
caverlee@cc.gatech.edu

Abstract

As organizations move toward using web technology to share data, the barriers go beyond the well known challenges of semantic interoperability. Our work proposes and illustrates the problem of establishing and maintaining data supply chains in the face of evolution and decentralized control. We explore agreements for *data* services, identifying obligations of provider and consumer, and opportunities beyond those for arbitrary services. Examples illustrate the kinds of capabilities that would be desired. We propose several open problems about data service agreements, e.g., techniques for agreement monitoring, and appropriate formalisms for specifying and reasoning about data supply chain graphs.

1. Introduction

Many research and industrial efforts (e.g., publish/ subscribe information brokers [EFGK03], web services, the semantic web) aim to provide looser coupling between information providers and consumers. Promised benefits include greater flexibility and the ability for consumers to rapidly discover and exploit new information sources.

Achieving this more flexible coupling requires addressing three challenges. First, semantic interoperability requires continued attention, despite the many results so far. Second, we must ensure that needed data is actually collected. That is, beyond integrating existing data, we must guide semantic choices in both new systems and new ontologies used to describe systems [RSR04].

The third challenge is the subject of this paper: how does one manage on-going data sharing relationships in the presence of evolution and decentralized control? As consumers increasingly build value-added services on top of data resources they do not control, we need tools to assist with configuration management in environments that lack centralized authority. The current state-of-the-practice is inadequate, with few options between the extremes of rigid configuration management boards with long approval and change cycles, and informal, ad hoc agreements with no supporting technical infrastructure.

The problem is analogous to supply chains for manufacturing organizations. For example, an automobile manufacturer wants not only access to their suppliers' stocks of parts; they want assurances that the parts will be available as needed throughout the manufacturing process. In contrast, current information brokers focus on matching current information needs and available data without supporting ongoing information supply chains.

Figure 1 shows an example, in which a military logistics organization wants daily updates over the next 3 months of the quantity of fuel at the depots of coalition partners A and B. The example extends the third part of the web services Publish-Find-Bind paradigm [GGKS02]. Publishers describe their offerings, and then consumers find services that meet their needs. Well known mediation

**Proceedings of the 30th VLDB Conference,
Toronto, Canada, 2004**

* Work performed while at The MITRE Corporation.

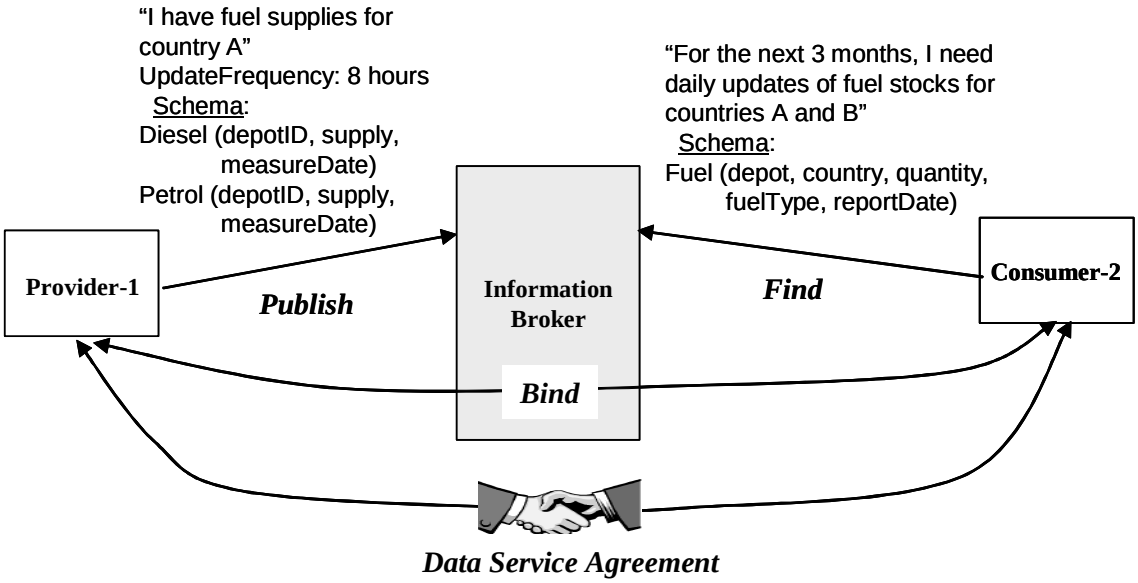

Figure 1. Logistics Example

techniques can identify matches, and recent research results (e.g., [GATM04]) help create bindings—e.g., create and potentially compile mappings between Provider-1's offerings and Consumer-2's information need. However, these techniques do not help manage the continuity of data flows in the face of evolution. Suppose Provider-1 has several requests for changes (e.g., to their schema, to the units that quantities are reported in, to collect data every 72 hours instead of every 8) and many consumers of their information. What changes affect which of their many consumers, and what are their responsibilities to each? Similarly, what are the consumer's obligations – authorized purposes, protection requirements, payment obligations? To address these needs, we introduce a new construct, the *data service agreement (DSA)*, and describe requirements for supporting tools, including:

- *Agreement specification*: to specify agreements between data providers[1] and consumers. These capture the obligations in a formal language that supports reasoning by software.
- *Violation detection*: to automatically detect certain violations to existing agreements
- *Notification*: to notify affected parties about changes in agreements and detected violations
- *Change impact analysis*.

Such tools would provide benefits by clarifying each participant's obligations in a data exchange. There are also benefits to specific types of participant:

- Data consumers, by increasing their confidence that they will continue to get the data they need
- Data providers, who gain confidence about usage, protections and payments
- Enterprise architects and planners, by giving them tools to analyze *data supply chains* (i.e., graphs of data service agreements). Analyses may help them determine vulnerabilities and the business value of different information systems.

Data service agreements aim to provide flexible but nonzero coupling between data providers and consumers, to enable important, continuing applications. The sections below describe data service agreements in greater depth, present an example, and discuss open research problems.

2. Data Service Agreements

A data service agreement specifies obligations and expectations of both the provider and consumer of a data service—i.e., a service which provides data from the provider to the consumer. The agreement specifies one or more of the following obligation types:

[1] We use "provider" and "supplier" interchangably. "Supplier" fits with the supply chain metaphor, while "provider" is the term used for web services.

- *Provider data service obligations*: these describe a provider's obligation to provide particular data to a consumer in accordance with specified quality and temporal constraints. The dual of this obligation is the consumer's *data expectation*.[2]
- *Consumer obligations*: these include *data protection*, *data usage*, and *compensation* obligations, described in Section 2.2.

Associated with each obligation is an action to be taken when the obligation is not fulfilled. Options include financial penalties and notifying someone (e.g., a DBA or a vice-president to whom the information flow is important).

DSAs are a specialization of service level agreements (SLA) between service providers and consumers [Wust02]. Typically, SLAs emphasize performance metrics about time (e.g., response or problem resolution time) or availability (e.g., maximum percentage of downtime). However, we are aware of no prior work on SLAs to address data obligations (e.g., the need to provide data of a certain quality at specified time intervals).

We now distinguish data services from arbitrary ones and then DSAs from SLAs.

A *data service* is one that performs conventional data operations (create, read, update, delete, a bit more), in terms of some set of physical or virtual data objects. This contrasts with general services, which can execute arbitrary code (e.g., HireEmployee). Data service agreements can exploit properties unique to data services, stemming from data algebras and operators. These properties include: (1) rules for combining or deriving data services, obtained from familiar (and easily analyzed) algebraic operations on the underlying data objects, (2) there are generic services for data (e.g., Read, Update, Notify), and known relationships among them; it is not necessary to re-establish them for each data object, and (3) there are well understood techniques for identifying overlap and subsumption relationships between consumer requests and views that describe provider offerings.

Compared with arbitrary services, for data we may be able to offer:

- *A natural clustering and taxonomy of agreements that affect portions of a data object.* For example, one agreement might concern particular columns; another might govern rows selected based on attribute values.
- *When a new information requirement is identified, one can examine whether the information involved is derivable from existing agreements that impose appropriate conditions.*
    - One may phrase some agreements in terms of information present, not a specific data structure. Conversion might be the consumer's responsibility (or the producer's or a middleman's). This is especially likely if the consumer's first action is to have a human vet the data.
- *Protection requirements on results.* Data is easier to pass on than services (e.g., can be emailed), and release is permanent – one cannot subsequently un-release it. Hence, release protections are very important. (Cryptographic techniques exist, but are problematic for data sharing and query [OSC03].) Also, privacy laws constrain data, not general services.

Figure 2 shows two data sharing agreements: between participants A and B and also between B and C. (The bold arcs show the direction of data flow.) DSA A_B describes what data A has promised (i.e., the result set of query Q1) and a quality parameter (i.e., the data—which contains satellite imagery—shall have resolution of 10 meters or better). The update restriction specifies the conditions that must be satisfied in order to change or delete the DSA, in this case that there must be signoff from a person in the role of the deputy director of organization N33. The notification requirement indicates that B must get at least 60 days notice before the change to the agreement can be instituted.

Agreement B_C promises C the result set of Q2 and that the data provided shall be no more than 20 minutes old. This agreement imposes no restrictions on B's ability to make changes to the agreement except that C must get at least 7 days advance notice. Finally, for B's obligation to C to hold, there is a precondition that DSA A_B is satisfied—i.e., B will be unable to provide the required data and meet specified quality parameters unless A satisfies his obligation to provide data to B.

Figure 2 also illustrates (on the lower right) some of the DSA services we envision. As shown, DSA services have interactions with the publish/subscribe information broker and metadata services.

2.1 Provider Data Service Obligations

The following shows our preliminary design of the structure of a data service obligation:

- What: i.e., what the provider is obligated to provide
    - A query defines the promised data[3]
    - Modality –e.g., provide full query result, provide the deltas, provide a continuous stream
    - Constraints (quality, recency and other constraints)
- When – once, at a list of times, periodically (with frequency specified)
- Preconditions[4] – defines the conditions under which the obligation applies, e.g.,

[2] The value of documenting consumers' semantic expectations was demonstrated by [GBMS99].

[3] We assume that other components mediate semantic discrepancies between sources and consumer needs.

    - StartDate, EndDate
    - CloudCover ≤ 30%
    - DataObligation-27 is satisfied (i.e., in order to provide this data, I must first receive the inputs I need to produce it)
- Information delivery mechanism (e.g., XML message, available at specified URL)
- Quality of service (QoS) parameters
- Update restriction – the conditions that must be satisfied in order to change or delete the data service obligation
- ChangeNotification – who must be notified, how to notify them, and the advance warning required when the Obligation is changed or deleted
- Violation actions – what will be done in the event that the obligation cannot be met.

2.2 Consumer Obligations

These are of three types:

- *Data protection obligation* – the consumer's obligation to protect the information against unauthorized release, in accordance with specified constraints (e.g., do not release beyond company X or the partners of consortium Y, protect the anonymity of research subjects, do not share with foreign governments, keep it on a machine well secured from external attack). Support for data protection obligations can leverage the literature on system protection and data releasability (e.g., [CFJF02]).
- *Data usage obligation* – the consumer's obligation to use the data only for specified purposes—e.g., for counter terrorism investigations, but not for ordinary criminal or tax investigations. While automated systems have little control over what use is made of information once it is acquired, there is value in documenting the obligations. In addition, one can relate groups of users to a specific purpose (e.g., tax investigation), and employ role-based access controls [PSA01].
- *Compensation obligation* – the consumer's obligation to compensate the provider for information provided—e.g., money, or a promise to provide positive feedback to management or to a reputation management system (as in eBay, Yahoo, Amazon, etc.).

As in Section 2.1, violation actions can be specified.

[4] The intent is to capture "hedges," conditions under which the obligation does not apply (e.g., satellite imagery will not be updated given cloudy conditions). There must also be a mechanism to express backup plans (e.g., if it is cloudy, we'll provide a different kind of imagery).

3. Research Issues

MITRE recently initiated a research effort to define, implement, and evaluate a technical infrastructure and usage methodology for data service agreements. We now present open research problems.

First, as shown in Figure 2, automated violation detection requires agents that monitor various system components. What kinds of automated detection are possible with different levels of intrusiveness? For example, what automated violation detection is possible by simply making the component systems' schemas queryable? One could also have the component systems implement log-based change notification. Considerably more intrusive would be requiring that the monitoring agents have permission to install triggers on component systems.

Second, since DSAs can include preconditions that other DSAs are satisfied, there are data supply chain graphs, which enable different kinds of analysis—e.g., determining what information resources are most critical to the operation of an enterprise. We will explore the properties of these graphs and what kinds of analyses are most useful.

Third, what formalism should be used to express agreements? Do binary agreements give enough power to capture most practical cases (in a manager's wizard and/or an internal form suitable for a reasoning engine)? Is the set of agreement constructs closed under composition? The Web Ontology Language (OWL) [SWM04] is attractive, because its description logic enables reasoning about DSA graphs. In addition, this would enable us to analyze and monitor systems that include semantic mediation—e.g., allowing us to notify an affected party using his view of the data.

Fourth, one might provide techniques to adjust the granularity of DSAs. Negotiation may be very labor intensive, and perhaps one does not want to trace every detail. Locking systems trade up (e.g., from record to page locks) and perhaps something analogous is appropriate here.

Finally, there is a need for pragmatic studies. How useful will DSA services turn out to be in practice? What technical infrastructure and guidelines will real organizations need to apply them?

References

[CFJF02] S. Chapin, D. Faatz, S. Jajodia, A. Fayad, "Consistent policy enforcement in distributed systems using mobile policies," *Data & Knowledge Engineering*, 43(3), 2002

[EFGK03] P. Eugster, P. Felber , R. Guerraoui, and A.-M. Kermarrec "The Many Faces of Publish/Subscribe", *ACM Computing Surveys*, Vol 35, No 2, June 2003

[GATM04] A. Gal, A. Anaby-Tavor, A. Trombetta, and D. Montesi, “A Framework for Modeling and Evaluating Automatic Semantic Reconciliation,” to appear in *VLDB Journal*, 2004

[GBMS99] C. H. Goh, S. Bressan, S. Madnick, M. Siegel: Context Interchange: New Features and Formalisms for the Intelligent Integration of Information. *ACM Trans. Inf. Systems*, 17(3), 1999

[GGKS02] K. Gottschalk, S. Graham, H. Kreger, and J. Snell, “Introduction to Web services architecture,” *IBM Systems Journal*, 41(2), 2002

[LKDK03] H. Ludwig, A. Keller, A. Dan, R. P. King, and R. Franck, “Web Service Level Agreement (WSLA) Language Specification”, http://www.research.ibm.com/wsla/, Jan 2003

[OASIS02] OASIS ebXML Collaboration Protocol Profile and Agreement Technical Committee, “Collaboration-Protocol Profile and Agreement Specification Version 2.0”, http://www.ebxml.org/specs/ebcpp-2.0.pdf, September, 2002

[OSC03] G. Ozsoyoglu, D. Singer, S.S. Chung, “Anti-Tamper Databases: Querying Encrypted Databases”, *IFIP WG 11.3 Working Conference on Database and Applications Security*, 2003.

[PSA01] J. Park, R. Sandhu and G. Ahn, “Role-Based Access Control on the Web,” *ACM Transactions on Information and Systems Security (TISSEC)*, 4(1), February 2001

[RSR04] A. Rosenthal, L. Seligman, S. Renner, “From Semantic Integration to Semantics Management: Case Studies and a Way Forward,” to appear, *SIGMOD Record*, September, 2004

[SWM04] M. K. Smith, C. Welty, and D. L. McGuinness, eds. “OWL Web Ontology Language Guide”, http://www.w3.org/TR/owl-guide/, Feb 2004

[Wust02] E. Wustenhoff, *Service Level Agreement in the Data Center*, Sun Microsystems, http://www.sun.com/blueprints/0402/sla.pdf, April 2002

Acknowlegement

The authors thank Scott Renner who provided the initial insight on the importance of data service agreements. We also thank Vipin Swarup and Ken Smith for helpful comments.

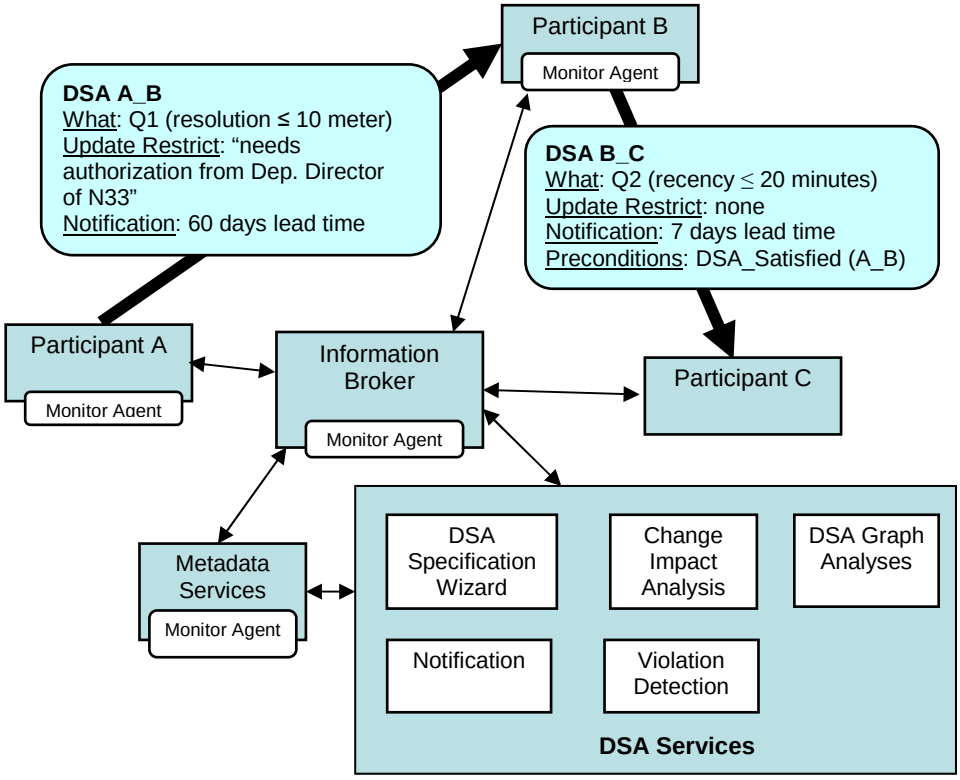

Figure 2. Example Data Service Agreements