A Lightweight Data Integration Architecture using Atom

David W. Williamson, Nigel J. Stanger
Department of Information Science, University of Otago
PO Box 56
Dunedin, New Zealand
+64-3-479-8142
{dwilliamson,nstanger}@infoscience.otago.ac.nz

ABSTRACT
Cost is a major obstacle to the adoption of large-scale data integration solutions by small to medium enterprises (SME's). We therefore propose a lightweight data integration architecture built around the Atom XML syndication format, which may provide a cost-effective alternative technology for SME's to facilitate data integration, compared to expensive enterprise grade systems. The paper discusses the underlying principles and motivation for the architecture, the structure of the architecture itself, and our research goals.

Categories and Subject Descriptors
K.4.4 [**Computers and Society**]: Electronic Commerce—*electronic data interchange (EDI)*; H.3.5 [**Information Storage and Retrieval**]: Online Information Services—*data sharing, web-based services*; H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*distributed systems*.

General Terms
Design, Economics, Experimentation, Measurement.

Keywords
data integration, Atom, SME, lightweight architecture, Semantic Web, B2B

1. INTRODUCTION
The ability to integrate data from multiple heterogeneous sources is becoming a key issue for modern businesses, and yet the number of businesses implementing data integration solutions is smaller than we might expect [2,20]. This is particularly true for small to medium enterprises (SME's), for whom the cost of implementing an enterprise-scale data integration solution can often be prohibitive [2,8,18].

In this paper, we propose a lightweight data integration architecture based on the Atom XML syndication format, which may provide a cost-effective alternative technology for SME's to facilitate data integration rather than having to purchase expensive enterprise grade systems. We are currently implementing a basic proof of concept of this architecture, and plan to evaluate it using three case studies.

The body of this paper comprises three main sections. In Section 2 we provide some general background information regarding data integration and the Atom syndication format. In Section 3 we discuss the motivation behind our proposed architecture. We then discuss the proposed architecture and the goals of our research in Section 4, and present some possible directions for future work in Section 5. The paper concludes in Section 6.

Submitted to *WebDB 2005*, June 16–17, 2005,
Baltimore, Maryland, U.S.A.

2. BACKGROUND
In this section, we briefly discuss the concepts and technologies that underlie our proposed architecture. In Section 2.1 we provide a brief overview of data integration, especially in the context of SME's attempting to implement a data integration solution. This is followed by a brief discussion of the development of Atom and related technologies such as RSS and RDF.

2.1 Data Integration
Data integration is a term used to describe the combining of data residing in different sources to provide the user with a unified view of data [1,22]. This activity is becoming increasingly important to modern business operation as more and more organizations rely upon applications that support staff in undertaking informed decision making [6,22].

Data integration is a domain that has been a topic of research for some time [2,21]; today this domain is of no less significance with many organizations requiring the aggregation of data from multiple and often heterogeneous sources, for a wide variety of applications [9]. Batini et. al. [1] illustrated three common scenarios for integration environments:

> *homogeneous*, where all the sources of data share the same schema;
>
> *heterogeneous*, where data must be integrated from sources that may use different schemas or platforms (e.g., a combination of relational and hierarchical databases); and
>
> *federated*, where integration is facilitated by the use of a common export schema over all data sources.

A typical example of data integration from heterogeneous sources can be found in the arena of business-to-business (B2B) commerce, where, for example, a manufacturer may have to interact with multiple suppliers or temporary contractors each of whom may have completely different data structures and data exchange formats [19]. With the introduction of cheaper web based technology, many additional organizations have been able

to undertake projects to facilitate data integration, however, the costs associated with such technology are still quite prohibitive to the many smaller companies and organizations that comprise the majority of most countries' economies.

Many initiatives have been put forward to try and alleviate this situation, one of the more recent being the OASIS Universal Business Language (UBL) standard [14], which is a project to standardize common business documentation—invoices, purchase orders etc.—so that it is easier for companies to establish and maintain automated transactions with other parties. UBL has been designed to operate with ebXML.

XML has been widely adopted as a standard platform for exchanging data between organizations, and many specialist standards—such as the aforementioned ebXML—have been developed to cater to the unique needs certain business sectors present. In addition to XML-based language specifications, other standards such as EDIFACT[1] and EXPRESS have been defined to facilitate the transmission of information from various sources so that it may be integrated with other data.

2.2 The Atom Syndication Format

In this section we provide a brief overview of the Atom syndication format and the technologies that led to its development.

2.2.1 RDF, RSS and the Semantic Web

The World Wide Web (WWW) as it stands today consists mostly of documents intended for humans to read, i.e., "*…a medium of documents for people rather than for data and information that can be processed automatically…*" [5], which provides minimal opportunity for computers to perform additional interpretation or processing on them [3,5]. In essence, computers in use on the Web today are primarily concerned with the parsing of elementary layout information, for example headers, graphics or text and processing like user input forms [4,5].

There are few means by which computers can perform more powerful processing or manipulation on web resources [5,7], most often because the additional semantics required do not exist or are not in a form that can be interpreted by computers [11]. The motivation for the adoption of semantics in Web documents can be made evident simply by using a contemporary search engine to look for an "address". This search may well return a plethora of results ranging from street addresses and email addresses to public addresses made by important individuals through the ages.

This kind of scenario is one of the reasons for the W3C's Semantic Web project [11]. In the words of its creator, Tim Berners-Lee, its goal is to:

> "*...develop enabling standards and technologies designed to help machines understand more information on the Web so that they can support richer discovery, data integration, navigation, and automation of tasks. With Semantic Web we not only receive more exact results when searching for information, but also know when we can integrate information from different sources, know what information to compare, and can provide all kinds of automated services in different domains from future home and digital libraries to electronic business and health services.*" [11]

In other words, the Semantic Web will provide a space where more intelligent searching and processing of information will be made possible by further extending the existing capabilities of the World Wide Web (WWW).

RDF is a technology that is an integral part of the W3C Semantic Web initiative, as the following excerpt from the W3C Semantic Web activity statement will attest:

> *"The Resource Description Framework (RDF) is a language designed to support the Semantic Web, in much the same way that HTML is the language that helped initiate the original Web. RDF is a frame work for supporting resource description, or metadata (data about data), for the Web. RDF provides common structure that can be used for interoperable XML data exchange."* [17]

What RDF does in the context of the Semantic Web is to provide the capability of recording data in a way that can be interpreted easily by machines, which in turn provides an avenue to *"...more efficient and sophisticated data interchange, searching, cataloguing, navigation, classification and so on…"* [17].

Since its inception in the late 1990's, the RDF specification has spawned several applications, RSS being but one example. RDF Site Summary (RSS) is an XML application, of which versions 0.9 and 1.0 conform to the W3C's RDF specification. It is a format intended for metadata description and content syndication [12]. Originally developed by Netscape as a means to syndicate content from multiple sources onto one page [16], RSS has been embraced by other individuals and organizations resulting in the spawning of multiple versions.

At its most simple, the information provided in an RSS document comprises the description of a "channel" (that could be on a specific topic such as current events, sport or the weather, etc.) consisting of URL linked items. Each item consists of a title, a link to the actual content and a brief description or abstract.

Because of the proliferation of differing RSS standards and associated problems with compatibility, a group of service providers, vendors and developers have initiated the development of a separate syndication standard named Atom, which will, according to the Atom Publishing Format and Protocol (Atompub) Working Group, be heavily influenced by the lessons learned in the evolution of RSS.

2.2.2 Atom

The Atom[2] specification is an XML-based document format that has been designed to describe lists of related information [16]. These lists are known as "feeds". Feeds are made up of multiple items, known as "entries"; each entry can have an extensible set of attached metadata [16].

Atom as a technology comprises four key related components: a conceptual model of a resource, a well defined syntax for this model, the actual atom feed format itself and the editing protocol.

[1] Further information on EDIFACT is available at http://www.unece.org/trade/untdid/welcome.htm.

[2] Atom information is available at http://www.atomenabled.org/.

Both the feed format and editing protocol also make use of the aforementioned syntax.

In addition to these features, the Atompub Working Group have outlined several design objectives for the feed format and the editing protocol. The feed format must be able to represent the following: a resource that is a weblog entry or article, a feed or channel of entries, a complete archive of all entries within a feed, existing well formed XML (especially XHTML) content and additional information in a user-extensible manner.

The editing protocol must support creating, deleting or editing feed entries, multiple authors for a single feed, user authentication, user management and the ability to create, obtain and configure complementary material such as comments or templates.

The latest specification of Atom, which at the time of writing is still in a draft form, states the main purpose that Atom is intended to address is *"...the syndication of Web content such as Weblogs and news headlines to Web sites as well as directly to user agents*" [16]. The specification also suggests that Atom should not be limited to just web based content syndication but in fact may be adapted for other uses or content types. The Atompub Working Group aim to submit the Atom feed format and editing protocol to the IETF for consideration as a proposed standard in early April 2005.

3. MOTIVATION

One of the example domains of data integration is that of Electronic Data Interchange (EDI), a concept used by companies to exchange information such as goods procurement documentation. EDI is not new [2,15], and has been used for many years by various organizations to reduce costs by replacing more traditional paper based systems. It is interesting to note, however, that in surveys regarding the extent of adoption of EDI, only a fraction of the companies that might be perceived as beneficiaries of such technology have actually implemented or attempted to implement it [2,20]. This naturally raises the question of why? We can refine this question further by asking why so few smaller companies (SME's) have adopted EDI or indeed other technologies that rely on accurate automated data integration, such as data warehousing.

Perhaps the most important reason is that of cost: to a small company the perceived benefits of introducing the technology may not be sufficient to justify the expense [2,8,18]. When a decision has been made to implement new technology, it is often the case that the SME in question has been forced into an investment that is, to them, an expensive solution, perhaps due to demands imposed by larger clients and partners, or as a response to competitors in an attempt to maintain market position [2,20].

Attempts have been made to make EDI more cost effective by introducing EDI on a web-based platform [2], and through the development of standards such as the recently sanctioned OASIS Universal Business Language (UBL) standard [14]. While UBL is new and has probably not had sufficient time to make a substantial impact, the fact remains that the underlying reason these types of technologies are still not attractive enough to SME's is cost [2,8,18,20].

To summarize, data integration related technologies are often not readily or willingly implemented by SME's because of the perceived high costs involved, and at best are implemented only if it is deemed vitally important to the continued survival of the organization in the marketplace.

Such a situation leads us to the conclusion that there is an apparent need for an alternative data integration solution that is cost effective, enabling SME's to embrace the benefits of applications that use data integration technologies, such as data warehousing, EDI networks or e-catalogues.

This identified need provides the motivation for our proposed architecture, which we will discuss in the next section.

4. PROPOSED ARCHITECTURE AND RESEARCH GOALS

To address the issue of lack of SME adoption of data integration technologies, we propose a lightweight data integration architecture based on Atom, as illustrated in Figure 1. Atom was chosen as the underlying technology because of its XML heritage, and because the Atom community is trying to encourage different uses for the format beyond the traditional application of weblog syndication [16]. Although the standard has yet to be officially ratified, it already has a large user and development community.

We are currently implementing a basic proof of concept of this architecture, and will evaluate its cost-effectiveness and performance compared to other data integration technologies. The prototype builds upon existing software available for processing Atom feeds, and adds a module (written in PHP) for integrating incoming data from different feeds.

The integration module takes as input Atom feeds from multiple data sources, which simulate incoming data from client or supplier data sets. (For the initial prototype we have assumed that the data feeds are homogeneous; obviously this will need to be extended to heterogeneous feeds in later versions.) After the Atom feeds have been collected, the integration module will integrate the data supplied by the feeds into a schema that matches that of the target database, as shown in Figure 1. A transaction simulator will be employed to simulate workload and updates to the source databases, in order to recreate a day-to-day production environment.

In order to evaluate the prototype, we will implement three different simulated scenarios derived from actual use cases of previous projects. All three case studies follow a similar structure whereby data will be exported as Atom feeds from the source database(s), which are then consumed by the integration module before being sent to the target database for insertion.

The first scenario will simulate the integration of product data from multiple suppliers into a vendor's product information database. The product information database is used to populate the vendor's online product catalogue, which clients use to make decisions regarding goods procurement. The Atom feeds in this scenario represent flows of product data from the supplier to the vendor.

The second scenario follows on from an earlier research project to develop a kiosk system for the sale and distribution of music in digital format. The database the kiosk(s) use will be populated with information from vendors who have agreed to supply content (e.g., a record label's collection of music files). What is needed is a mechanism to integrate all the music data from each supplier into the music kiosk system's own database. The Atom feeds in this scenario are used to maintain an up to date database that has

the location and description of each available music track for sale in the system.

The third scenario will simulate the implementation of a data warehousing solution for a computer components distributor.

Preliminary results from the case study evaluations are expected to be available by June 2005. Our primary goal with the initial prototype is to prove the feasibility of our approach. We will compare our proposed architecture against existing data integration solutions by means of a cost/benefit analysis. We may also investigate measuring various software quality characteristics as defined by the ISO 9126 standard [10].

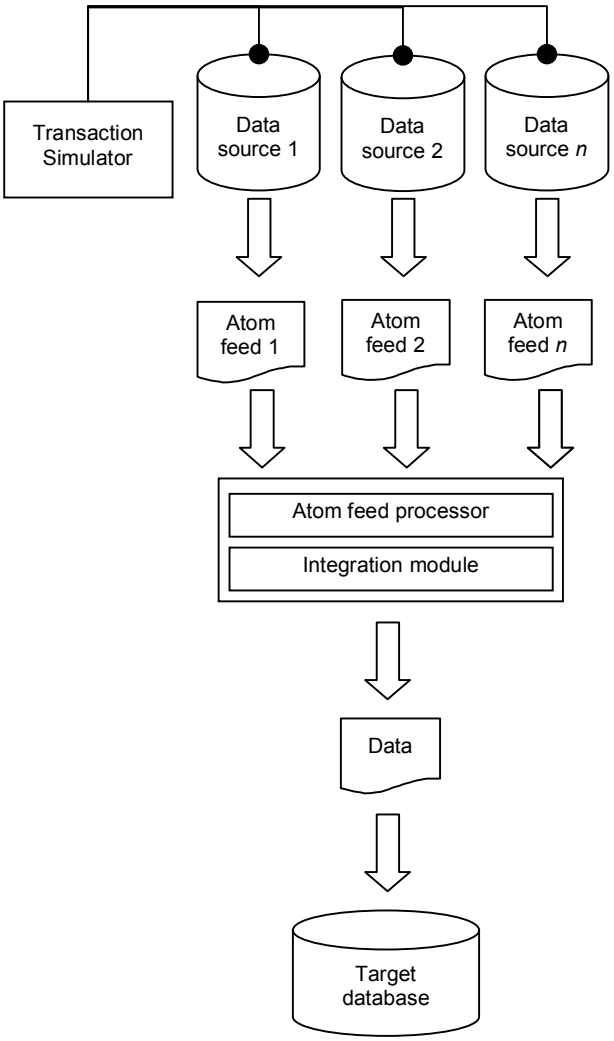

Figure 1. Proposed architecture showing integration module

5. FUTURE WORK

As the initial prototype is intended as a basic proof of concept of our proposed architecture, it has been kept as simple as possible in order to facilitate the implementation and evaluation. There are several obvious extensions to the basic prototype that will be investigated in later iterations of the architecture.

The initial prototype assumes that all data sources are largely homogeneous, that is, that they all share similar semantics and can therefore be relatively easily integrated. An obvious extension is to permit heterogeneous data sources that have differing semantics. Such an extension would require the addition of an ontology management module between the Atom feed processor and the integration module. This module will probably be based around the W3C's Web Ontology Language (OWL) [13].

The initial prototype also assumes only a single "author" per Atom feed, that is, there is only a single database underlying each feed (as implied by Figure 1). We can envisage a situation where what appears to be a single data source is actually a view layered on top of a collection of underlying databases (e.g., a supplier might draw data for their Atom feed from multiple databases within their organization). It would therefore be useful to investigate the possibility of multiple "authors" per Atom feed. This could imply an additional layer of data integration within the data source itself.

The data flows shown in Figure 1 imply that the proposed architecture is one-way only (i.e., from the data sources to the target database), but this may not be true in general. It would therefore be interesting to investigate extending the architecture to allow for the possibility of two-way data transfers, i.e., allowing data to flow from the target back to the sources.

6. CONCLUSION

In this paper, we discussed a lightweight data integration architecture based on the Atom XML syndication format. Cost is a major factor in the slow adoption of data integration technologies by small to medium enterprises, so the proposed architecture could provide a cost-effective alternative for implementing data integration infrastructures in small business environments. We are currently developing a basic proof-of-concept prototype system that will be evaluated using a series of realistic case studies. We expect to have preliminary results from these evaluations by June 2005.

7. ACKNOWLEDGMENTS

The authors would like to thank Dr. Colin Aldridge and Dr. Stephen Cranefield for their helpful comments on an early draft of this paper.

8. REFERENCES

[1] Batini, C., Lenzerini, M., and Navathe, S. B. (1987). A comparative analysis of methodologies for database schema integration. *ACM Computing Surveys*, *18*, 4 (Dec. 1986), 323–364.

[2] Beck, R., Weitzel, T., and König, W. (2002). Promises and pitfalls of SME integration. In *Proceedings of the 15th Bled Electronic Commerce Conference* (Bled, Slovenia, June 17–19, 2002). 2002.

[3] Berners-Lee, T., and Fischetti, M. *Weaving the Web.* Orion Business, London, 1999.

[4] Berners-Lee, T., Connolly, D., and Swick, R. R. (1999) *Web Architecture: Describing and Exchanging Data.* W3C Note, World Wide Web Consortium, 7 June 1999. http://www.w3c.org/1999/04/WebData

[5] Berners-Lee, T., Hendler, J., and Lassila, O. The Semantic Web. *Scientific American*, *284*, 5 (May 2001), 34–43.

[6] Calvanese, D., De Giacomo, G., Lenzerini, M., Nardi, D., and Rosati, R. Information integration: Conceptual modeling and reasoning support. In *Proceedings of the 3rd IFCIS International Conference on Cooperative Information Systems (CoopIS'98)* (New York, NY, August 20–22, 1998). IEEE Computer Society Press, Los Alamitos, CA, 1998, 280–291.

[7] Fensel, D., Hendler, J., Lieberman, H., and Wahlster, W. (Eds.) *Spinning the Semantic Web*. MIT Press, Cambridge, MA, 2003.

[8] Guo, J., and Sun, C. Context representation, transformation and comparison of ad hoc product data exchange. In *Proceedings of the 2003 ACM Symposium on Document Engineering (DocEng '03*) (Grenoble, France, November 20–22, 2003). ACM Press, New York, NY, 2003, 121–130.

[9] Haas, L. M., Miller, R. J., Niswonger, B., Tork Roth, M., Schwarz, P. M., and Wimmers, E. L. Transforming heterogeneous data with database middleware: Beyond integration. IEEE Data Engineering Bulletin, 22, 1 (Mar. 1999), 31–36.

[10] ISO. *Software Engineering—Product Quality—Part 1: Quality Model*. Standard ISO/IEC 9126-1:2001, International Organization for Standardization, Geneva, Switzerland, 2001.

[11] Koivunen, M., and Miller, E. W3C Semantic Web activity. In *Semantic Web Kick-Off in Finland: Vision, Technologies, Research, and Applications* (Helsinki, Finland, November 2, 2001). HIIT Publications, Helsinki, Finland, 2002, 27–43.

[12] Manola, F., Miller, E., and McBride, B. *RDF Primer*. W3C Recommendation, World Wide Web Consortium, 10 February 2004. http://www.w3.org/TR/rdf-primer/

[13] McGuinness, D. L., and van Harmelen, F. *OWL Web Ontology Language: Overview*. W3C Recommendation, World Wide Web Consortium, 10 February 2004. http://www.w3.org/TR/2004/REC-owl-features-20040210/

[14] Meadows, B., and Seaburg, L. *Universal Business Language 1.0*. OASIS Committee Draft cd-UBL-1.0, Organization for the Advancement of Structured Information Standards, Billerica, MA, 15 September 2004 http://docs.oasis-open.org/ubl/cd-UBL-1.0/

[15] Medjahed, B., Benatallah, B., Bouguettaya, A., Ngu, H. H. A., and Elmagarmid, A. K. Business-to-business interactions: Issues and enabling technologies. *The VLDB Journal*, *12*, 1, (May 2003), 59–85.

[16] Nottingham, M., and Sayre, R. *The Atom Syndication Format*. IETF Internet-Draft draft-ietf-atompub-format-06, Internet Engineering Task Force, 12 March 2005. http://www.ietf.org/internet-drafts/draft-ietf-atompub-format-06.txt

[17] Powers, S. *Practical RDF*. O'Reilly & Associates, Sebastopol, CA, 2003.

[18] Sommer, R. A., Gulledge, T. R., and Bailey, D. The n-tier hub technology. *ACM SIGMOD Record*, *31*, 1 (Mar. 2002), 18–23.

[19] Stonebraker, M., and Hellerstien, J. M. Content integration for E-Business. In *Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data (SIGMOD '01)* (Santa Barbara, CA, May 21–24, 2001). ACM Press, New York, NY, 2001, 552–560.

[20] van Heck, E., and Ribbers, P. M. The adoption and impact of EDI in Dutch SME's. In *Proceedings of the 32nd Hawaii International Conference on System Sciences (HICSS-32)* (Maui, Hawaii, January 5–8, 1999). IEEE Computer Society Press, Los Alamitos, CA, 1999, 7061.

[21] Wiederhold, G. Intelligent integration of information. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data (SIGMOD '93)* (Washington, D. C., May 26–28, 1993). ACM Press, New York, NY, 1993, 434–437.

[22] Yu, C., and Popa, L. Constraint-based XML query rewriting for data integration. In *Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data (SIGMOD '04)* (Paris, France, June 13–18, 2004). ACM Press, New York, NY, 2004, 371–382.