

AN XML-BASED NETWORKING METHOD FOR CONNECTING DISTRIBUTED ANTHROPOMETRIC DATABASES[†]

H Cheng^{1*} and K Robinette²

^{*1} Human Effectiveness Directorate, Air Force Research Laboratory, 2800 Q Street, Wright-Patterson AFB, OH 45433-7947, USA, Email: huaining.cheng@wpafb.af.mil

² Human Effectiveness Directorate, Air Force Research Laboratory, 2800 Q Street, Wright-Patterson AFB, OH 45433-7947, USA, Email: kathleen.robinette@wpafb.af.mil

ABSTRACT

Anthropometric data are used by numerous types of organizations for health evaluation, ergonomics, apparel sizing, fitness training, and many other applications. Data have been collected and stored in electronic databases since at least the 1940s. These databases are owned by many organizations around the world. In addition, the anthropometric studies stored in these databases often employ different standards, terminology, procedures, or measurement sets. To promote the use and sharing of these databases, the World Engineering Anthropometry Resources (WEAR) group was formed and tasked with the integration and publishing of member resources. It is easy to see that organizing worldwide anthropometric data into a single database architecture could be a daunting and expensive undertaking. The challenges of WEAR integration reflect mainly in the areas of distributed and disparate data, different standards and formats, independent memberships, and limited development resources. Fortunately, XML schema and web services provide an alternative method for networking databases, referred to as the Loosely Coupled WEAR Integration. A standard XML schema can be defined and used as a type of Rosetta stone to translate the anthropometric data into a universal format, and a web services system can be set up to link the databases to one another. In this way, the originators of the data can keep their data locally along with their own data management system and user interface, but their data can be searched and accessed as part of the larger data network and even combined with the data of others. This paper will identify requirements for WEAR integration, review XML as the universal format, review different integration approaches, and propose a hybrid web services/data mart solution.

Keywords: Anthropometry, Database, Web Services, XML, Data Warehouse, WEAR

1 INTRODUCTION

Anthropometric data are typically collected from anthropometric surveys conducted by countries over their population samples. Each survey captures the anthropometric characteristics of the population at the survey time, and many surveys may be performed for the same population over years to record changing trends. With the advancement of technology and new engineering requirements, a modern anthropometric survey produces much more data than the traditional summary statistics for a population. It emphasizes not only the extraction of summary statistics but also the preservation of raw data so that special interests on a subset of the survey population can be satisfied. For example, the principal products of the Civilian American and European Surface Anthropometry Resource (CAESAR) survey (Robinette, Blackwell, Daanen, Fleming, Boehmer, & Brill et al., 2002) consist of the following raw data files and documentation (storage method shown in parentheses):

1. demographic data for each subject (database)
2. 3-D laser scan models for each of 3 postures for each subject (binary graphical files)
3. 3-D landmarks for two postures (binary graphical files)
4. traditional style measurements of three types (database)
5. summary statistics reports (text files)

In addition, the CAESAR survey was conducted through international cooperation that sampled populations in North America, The Netherlands, and Italy. Therefore, the type of data encompassed by the WEAR integration can be characterized as distributed, heterogeneous, over time, and large volume. The mission of the WEAR group is to integrate these resources and make them accessible and searchable to the public. Currently there are a dozen members in the WEAR group including government research institutes, academia, and private companies located around the world.

[†] Approved for public release; distribution is unlimited. AFRL-WS 06-2697 (Nov 21 2006)

During the last decade, the architecture for systems integration has evolved from fat-client type two-tier client-server to multi-tier component-based distributed internet architecture. The latter represents the de facto architecture for custom-developed enterprise solutions since the late 1990s. Components located on dedicated application servers share and manage pools of connections, communicate via APIs (Application Programming Interface) if residing on the same server or RPC protocol (Remote Procedure Call) if crossing server boundaries. The platform is generally efficient (open connection) and reliable (guaranteed execution) but very complex. The increasing sophistication of middleware from major vendors lessened the overall complexity somewhat. For this type of integration, the road maps and tools are readily available from vendors.

Component-based distributed architecture requires the use of proprietary communication protocols for RPC, such as DCOM (Distributed Common Object Model) from Microsoft or vendor/network-dependent implementation of CORBA (Common Object Request Broker Architecture). This may not be a problem for most enterprise systems integration because of the vertically integrated organizational structure. However, with the emergence of E-commerce, this architecture has difficulty in accomplishing business-to-business (B2B) integration among independent business partners who mostly do not want to adopt and adhere to a proprietary solution. The introduction of XML and web services technologies lays the foundation for the introduction of service-oriented architecture (SOA). SOA offers an alternative to the traditional architecture, especially for the B2B type of integration. It expands the traditional distributed internet architecture into an open and interoperable computing platform.

In the following section, this paper tries to identify the integration issues faced by the WEAR group, evaluate both the traditional and service-oriented architecture's ability in addressing these issues, and propose the best solution for WEAR integration.

2 WEAR INTEGRATION ISSUES

To find the best integration solution for the WEAR integration, the problems have to be analyzed in more detail to identify the issues needed to be solved by the solution - in other words, the criteria for defining a successful integration must be established.

Based on the WEAR data collections and its mission, the solution has to address:

1. Data extraction and search

This is the ability to search and retrieve data from distributed and disparate data sources. Because of the remote location of WEAR members, distributed data sources connected through public Internet are the norm for WEAR. Though most of the anthropometric data are in the structured repositories such as databases, there are large numbers of non-structured data such as standalone body scan files that may or may not be stored in a database.

2. Data representation

This is the ability to transform anthropometric data into a universal format that is meaningful to and can be adopted by most anthropometry researchers and users. Because WEAR consists of data resources from different locales, the data format should be able to encode character sets that are not ASCII, such as French, Korean, Japanese, and so on. In addition, considering the potential wide use of WEAR group data, this format has to be open source, reusable, extensible, and vendor-independent.

3. Data quality

This is the ability to cleanse and enforce integrity for data extracted from disparate sources. Most of the anthropometry data are collected from field studies involving large numbers of subjects. In addition, different surveys done over years tend to use different sets of measurements. When retrieving and forming datasets of multiple surveys from multiple members, some algorithms are needed to match and cleanse data before any use.

4. Data analysis

This is the ability to analyze integrated datasets and offer solutions to users' requests. These solutions can be developed into a set of toolkits and made accessible to various end users, such as anthropologists, ergonomics engineers, clothing designers, etc. Therefore, analytical capability should be part of the integration design criteria.

5. Network accessibility

This is the ability to access data residing in different computer networks. Anthropometry databases in the WEAR groups reside in computer networks ranging from private company and academia to government agencies and military organizations. Each of them may have different rules and restrictions for accessing data. A least restrictive way of passing data over firewalls of these networks is critical for the successful WEAR integration.

6. Security

This is the ability to secure data and perform user authentication and authorization. Even though the WEAR integration promotes the sharing of anthropometric data, future cooperation among more industrial groups means that the integration has to take into account the protection of potential proprietary or restricted datasets. Besides securing data, user authentication and authorization are needed to manage end users of WEAR data.

7. Data/application autonomy

This is the ability for each WEAR contributing member to maintain complete control of its resources and application interface, encapsulate its business logic, and have choice of its platforms. Autonomy also provides scalability. This is not a mere technical issue. Integration with a desirable level of autonomy can help to address some organizational challenges that may arise during the process. Lack of cooperation from relevant parties is one of the major reasons for failure of many data integration projects because data owners are usually reluctant to offer their data unconditionally. One way to alleviate this concern is to build enough autonomy into the integration architecture so that all data contributing members will buy whole-heartedly into the integration objectives and approaches.

8. Financial and manpower constraints

Because the WEAR group is a nonprofit organization, its operation and resources are supported by member contributions and outside sponsorships. There isn't a centralized pool of funds and developers dedicated to the WEAR integration, at least in the early stage of the effort. The beginning work has to be done by members according to their own budgets, and the integration mostly has to be made in a piece-by-piece and phase-by-phase style. Therefore the architecture has to be flexible and scalable enough to accommodate this type of development process.

3 XML IN WEAR INTEGRATION

Among the above integration issues, data representation is a fundamental one. Because XML has become the de facto standard for sharing data, it is natural for the WEAR group to use XML as the standard for data representation. An anthropometric XML document will have the following benefits:

1. Open standard and independent of platforms
2. Ability to represent data in a universal and validated format for cross-platform sharing
3. Ability to encode beyond ASCII character sets
4. Human and machine readable
5. Ability to pass over firewall through HTTP port
6. Extensible, so vocabulary can be increased later

These benefits help to address several of the above integration issues. In order to achieve this automatic exchange and sharing of anthropometric data using XML, the WEAR group has to develop and agree on an XML schema (an XML document by itself) to define the legal vocabulary of and hierarchical relationships in anthropometry XML documents. If all WEAR members publish data according to the WEAR XML schema, a universal format of data representation has been achieved.

With the XML schema, supporting tools or customized algorithms can be used to establish bi-directional mapping of XML schema and database tables. This mapping helps to load XML documents into database tables or package query results into XML documents. Most database platforms already have these tools. More recently, native XML data types are introduced in most database platforms but may not have much use in the WEAR integration because of already established database schemas within the WEAR group.

Each member of the WEAR group is responsible for the development of mapping between its platform and WEAR XML schema. Because of the structural difference between the XML's tree-type hierarchy and a relational database's relationship, the design of WEAR XML schema should avoid complex hierarchy and recursive reference. It should focus on the efficiency of data exchange rather than on the detailed modeling of anthropometric data. Otherwise, the mapping will be difficult to build. In addition, the design also needs to take into consideration the possible importation to other larger schemas, such as biomechanics and biometrics.

4 INTEGRATION SOLUTION OPTIONS

Over the years, there have been many solutions developed for integrating distributed data sources, especially in the area of enterprise systems integration and business-to-business (B2B) integration. This section evaluates several typical solutions and their WEAR applications in order to highlight the choice presented by this paper.

4.1 Linked Sever

Linked server architecture is straightforward conceptually. Figure 1 (Microsoft, 2006) depicts its basic configuration in Microsoft SQL Server. Linked server specifies an OLE DB provider (DLLs) to manage and interact with OLE DB data sources. The data sources can be databases or other file formats, such as text files and spreadsheets etc., as long as there are OLE DB providers available for the formats.

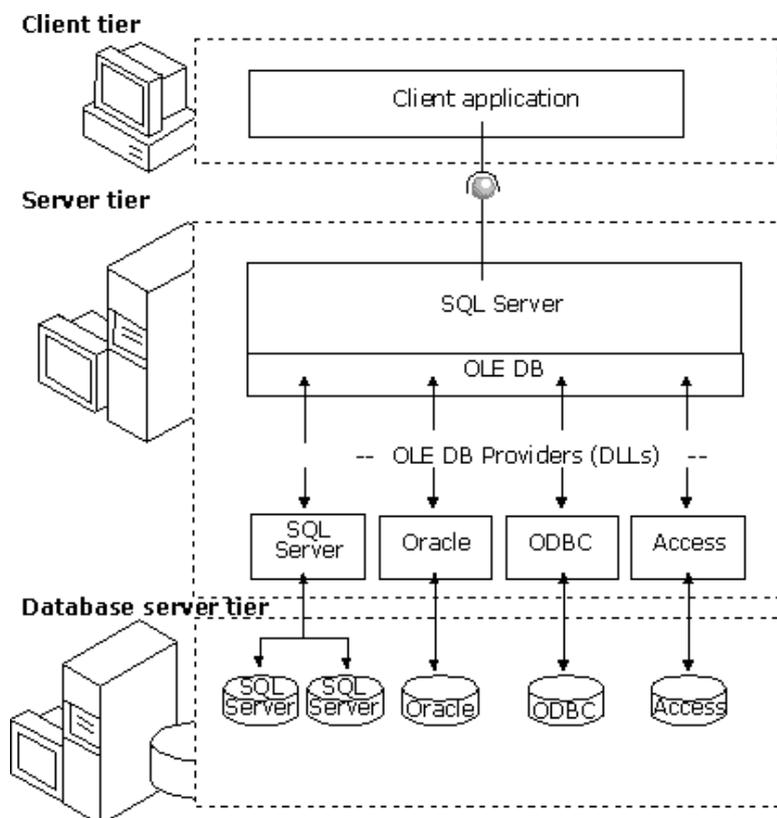


Figure 1. Linked server configuration (Microsoft, 2006)

The main advantage of linked server is that it allows running distributed queries, updates, commands, and transactions on distributed data sources. Very similar to the regular queries, the distributed queries use four-part name syntax involving the linked server name instead of only the table name. Distributed queries allow

exploring different data sources in a similar and efficient way. For differences in databases structures, views can be used to achieve some degree of similarity among data sources for the execution of distributed queries. Linked server is a mature technology, and the implementation is straightforward compared to the other solutions.

Even though the linked server has the advantage of simplicity in addressing the WEAR integration issues, it has the following drawbacks:

1. It is a maintenance and update nightmare because of the very tight integration. Any changes in the databases may easily break up the links. In addition, it cannot satisfy the data/application autonomy requirement and therefore is not very scalable.
2. It requires direct database access and reliable connection, which is difficult to accomplish within the WEAR group. The network access rules applied to some WEAR group members may deny any direct access to the database servers. Therefore, the linked server cannot satisfy the network accessibility requirement.

4.2 ETL and Data Warehouse

Integration process ETL (Extract, Transform, and Load) is usually combined with data warehouse as the solution for enterprise systems integration because the combination has the capabilities of not only data integration but also data analysis. Data warehouse is usually built on top of operational databases and other possible semi-structured or non-structured data sources for conducting business intelligence analysis. The underlying operational databases collect and update vast amounts of information in real time from many sources/users. Through an integration process such as ETL, these transactional data are loaded into the data warehouse for analysis of data over time, such as trend and variance analysis through OLAP (On-Line Analytical Processing) and data mining. Because WEAR data are not transactional data, ETL and data warehouse seem to fit as the solution for the WEAR integration, particularly the data warehouse which can provide great potential in the use of WEAR data.

Data warehouse is designed mostly for analytical purposes. It tends to load data once, queries data many times with complex queries, and rarely if ever changes data. The data need to be cleansed and validated before being loaded into the data warehouse. Because of these characteristics, data warehouse takes a highly normalized structure of many tables connected with a series of PK-FK (primary-foreign key). It ensures each item of data is stored only once (minimal redundancy) and provides a flexible means to define relationships. This highly normalized structure is efficient for adding/updating/storing data but is not intuitive for business analysis because of complex queries with many joins. Therefore, more intuitive structures such as de-normalized star schema and others are used to form data marts from the data warehouse, while the data warehouse is serving as the system-of-records. Data marts are subsets of information relevant to a particular area of interrogation.

The star schema is a multidimensional data model consisting of a central fact table and many dimension tables. The facts are numerical measures and aggregations. The dimensions are the entities with respect to which the facts are kept and analyzed. The fact and dimension tables are related through dimension keys. Unlike the commonly used entity-relationship data model of relational databases, the multidimensional model is a subject-oriented schema that facilitates data analysis instead of transaction process. For potential anthropometric data warehouses, this type of star schema and data mart may have some interesting applications.

For example, the following star schema shown in Figure 2 can be used to explore the possibility of finding groups of anthropometric measurements that can be used to distinguish different groups of people against dimensions of birth place, ethnicity, gender, and age. Anthropometric measurements can be analyzed and mined against any possible subsets of these given dimensions (cuboid). If a distinct pattern exists between certain measurements and cuboids, it can be used as a biometrics identifier or parameters for other research, such as social studies.

This application demonstrates that data warehouse as the integration and analysis model has some clear advantages. Because of data cleansing and validation through ETL processes, data loaded into the data warehouse possess the high quality that is required by OLAP and data mining algorithms. The knowledge and discovery from OLAP and data mining are the most valuable products one can get from data integration.

Nowadays enterprise integration using ETL and data warehouse is a mature technology supported by many vendor products. Tools such as SQL Server 2005 Integration Services (SSIS) are available to facilitate the

implementation of ETL processes against distributed and disparate data sources. They can integrate data sources that are distributed, structured or unstructured, large volume, and asynchronous.

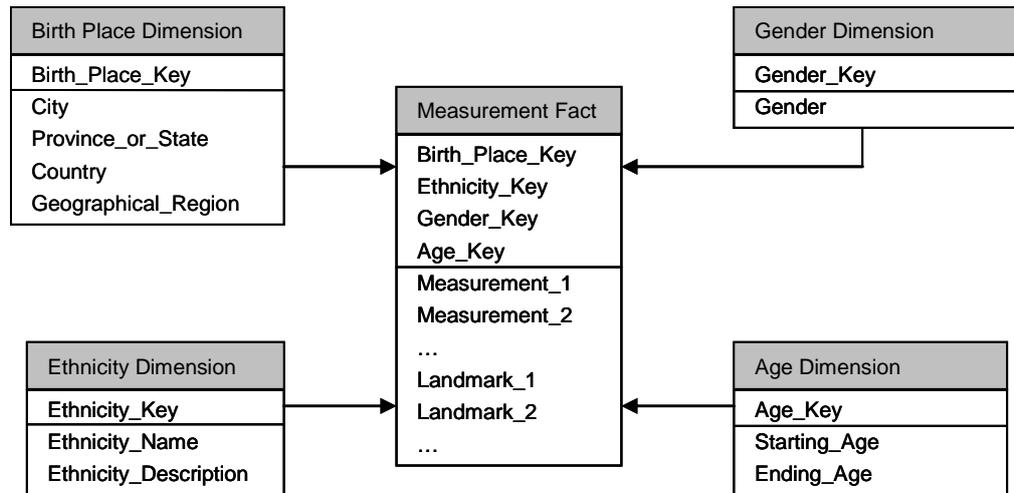


Figure 2. An example of a star schema model of anthropometric data

Even though tools such as SSIS offer comprehensive solutions, they are mostly products of individual vendors and their implementations are more or less platform-dependent. It is difficult for all WEAR members to agree on the use of vendor-dependent proprietary integration tools. In addition, the implementation of ETL processes often results in a tight integration that affects autonomy. It requires great effort to develop and coordinate ETL processes among independent WEAR members with various system platforms and database structures. Another downside of this integration is the accompanying high development and maintenance cost and special expertise.

Nevertheless, the analytical advantages brought by the data warehouse are very beneficial to WEAR, and it is worth exploring ways to incorporate data warehouse into the WEAR integration.

4.3 XML Web Service

During the last decade, IT enterprise architectures have evolved from client-server architecture to component-based distributed Internet architecture. The linked server and data warehouse are two uses of these structures and represent tight data integration. With the introduction of XML web service technology, service-oriented architecture (SOA) concept was introduced and gained much attention. Opposite to the tight integration, XML web service is a type of loose integration. Though the technology is still in a fast-evolving state, much progress has been made with the adoption of basic open standards such as XML and SOAP by major industry vendors. The maturity of the basic standards and comprehensive platform supports by major vendors present a new alternative approach to the WEAR integration.

The XML web service architecture adheres to the principles of service-orientation. According to the principles, services in SOA are reusable, loosely coupled, autonomous, stateless, and discoverable. They share a formal contract, abstract underlying logic, and may compose other services (Erl, 2006).

An application of XML web service architecture to the WEAR integration can be illustrated by Figure 3. The model is defined by three basic entities of web services: the service requestor, the service provider, and the service registry. The three entities operate according to the following set of web service standards:

- WSDL - It stands for Web Service Description Language, an XML document that describes a web service, how to call the service, and what to expect as a response from the service.
- SOAP – It is considered a name and not an acronym. It defines the XML format for messages (data and remote procedure call) sent from computer to computer, i.e., between the service provider and requestor.
- UDDI – It stands for Universal Discovery Description and Integration. It provides the standardized service registry. It is the “yellow page book” of web services.

Except for UDDI that is still an optional discovery mechanism, WSDL and SOAP, based on XML, have become core technologies of web services and SOA. They are adopted by most industry vendors.

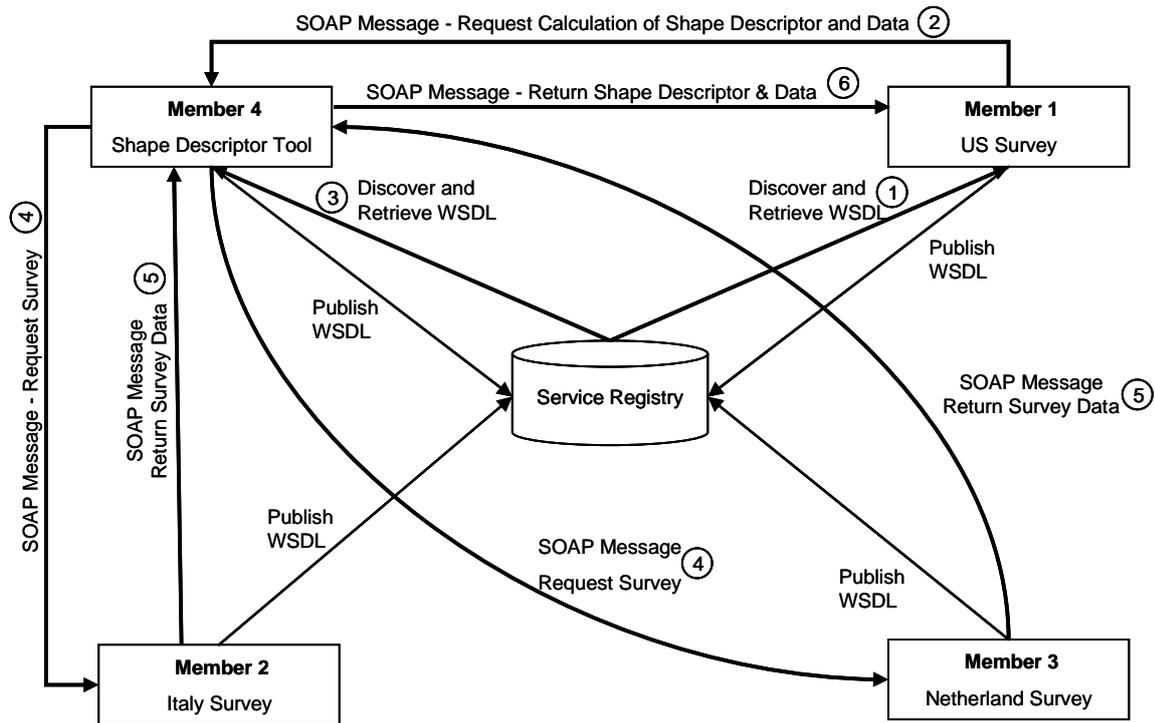


Figure 3. An example of XML web service of WEAR integration

Figure 3 demonstrates a hypothetical WEAR data integration using XML web service. The entities in this model are four members of the WEAR group. All of them publish their services to the service registry using WSDL. Three members (member 1, 2, 3) offer sharing of their survey data, i.e., Italy, The Netherlands, and US adult anthropometry surveys. Member 4 offers a unique tool which calculates shape descriptors from survey data. Shape descriptor is a 128-bit number used to search and group bodies that have similar body shapes. If member 1 needs a dataset consisting of certain subjects from Italy and The Netherlands surveys as well as their shape descriptors, it can obtain it from other members' services following the steps from 1 to 6. All the requests and return data will be wrapped in the bodies of SOAP messages and travel among members' web sites.

The loose integration of XML web service is achieved by offering functions and resources as services according to services contracts (WSDL) and requesting/delivering them as payloads in the SOAP messages through some transport protocol such as TCP/IP. This messaging mechanism is one of the web service foundations and it brings:

- Statelessness. Entities minimize the amount of state information they manage and the duration for which they hold it. The request and response have all the information necessary to understand them. The connection between entities is closed once the message is passed over.
- Autonomy. Entities have distinct boundaries and are independent. They encapsulate their logistics and underlying resources and have complete control over them. They offer functionalities and resources as services.

Therefore, XML web service can address the most significant drawbacks of tight integration - lack of autonomy and scalability. Because the message format is usually XML-based (regardless of SOAP or not), it can be constructed according to the strict XML schema agreed to by all the service providers and understood by the clients. The use of SOA and XML can address many WEAR integration issues – extraction and search, network accessibility, data transformation, data quality, data/application autonomy, scalability, and some security (SSL).

Though XML web service has unique advantages in meeting the challenge of WEAR integration, it does have a few weak areas. One of the areas is the user authentication and authorization, which is more of a challenge in the XML web service environment. For example, beyond the initial service request, most of the communication

is done by programs, and there is a need to propagate the service requestor's authentication and authorization information through multiple service providers. Basically for the user authentication and authorization, XML web service needs a standard for single sign-on as well as role-based access control. The second-generation web service specifications (WS-* specifications) are addressing these and other problems. However, it is still an ongoing process.

5 WEAR INTEGRATION SOLUTION

For the two types of integrations - enterprise and B2B - it is apparent that ETL and data warehouse are better for the former and XML web service is better for the latter. However, the WEAR integration does not fall exactly into either category. It is neither an enterprise integration because WEAR members operate independently nor a B2B because the integration will be achieved within the WEAR group. The integration can be more accurately defined as a federated systems integration. In order to find a workable solution, one can divide the objectives of WEAR integration into short-term and long-term ones.

The short-term objectives can be summarized as:

- Integrate and share anthropometry survey data automatically among the WEAR members
- WEAR members maintain and control independently their databases and existing web applications.

The implementation of the short-term objective will result in an integrated WEAR though it won't address the issues of data analysis. From the above evaluation, XML web service is the best solution to this short-term objective. Because of the loosely coupled nature, web service is inherently federated and allows each member the autonomy of deciding what to share. The use of XML allows producing universal anthropometric data sets from distributed and disparate data sources. The solution is also workable with the current state of web service technology, and the implementation can be further simplified by the following measures:

- Limit the services only to datasets without RPC.
- Use restricted UDDI registry that allows publishing and discovering only within the WEAR group members. This increases the security.
- Implement the user authentication using X509 digital client certificate. This is manageable within the WEAR group because of its limited number of members. This is also excellent for internet use, which is the case for the WEAR group's remotely located members. The user authorization can be handled using SOAP header.

Because anthropometric data is not transactional data, real-time performance and transaction control are not concerns for the integration. Overall, the first-generation web service standards (WSDL, SOAP, and UDDI) are enough for the integration solution. We can avoid the issue of support on the second-generation web service (WS-*). Since first-generation web service is well supported by major vendors, the development cost will be relatively low and the solution is workable.

The long-term objectives of the WEAR integration are to offer services to the general public and industry groups in the following forms:

- Customized and integrated anthropometric datasets
- Anthropometry solution toolkits to standard problems from researchers and engineers
- Commercial design process management such as clothing and vehicle compartment designs.

These objectives require building strong analytical models into WEAR. This can be achieved by incorporating data marts into WEAR. The data marts can be built on the top of WEAR integration, produced for the satisfaction of short-term objectives instead of traditional ETL and data warehouses. This is possible because universal XML data files validated against strict anthropometric XML schema are high-quality data. Therefore, ETL processes can be replaced by the XML web services of WEAR integration. Because of the complexity of the analytical models and high demands on computation efficiency, data marts and toolkits should be built and maintained centrally instead of being offered separately by different WEAR members. Since the target of these capabilities is the outside users, XML web services with public scope of UDDI registry can serve as the major delivery means. This new architecture can be termed as the hybrid web service/data mart model. It is illustrated in Figure 4 as the solution to the long-term objectives of WEAR. In this model, various data marts are grouped and made into a virtual member "A" of WEAR group. The toolkits developed from these data marts or datasets

extracted from WEAR are offered as web services by member “A”. Instead of being restricted to the WEAR group internally, these web services are published to the public UDDI registry and are available to the general public or specific outside customers depending on their account rights. Essentially member “A” is the public interface to the WEAR group’s data sources and analytical tools, while the web services of other physical WEAR members will serve as the data feeders to member “A”. It should be pointed out that this architecture does not prevent individual WEAR members from publishing their own web services to the general public by replicating the registration of their public accessible web services to the public UDDI registry.

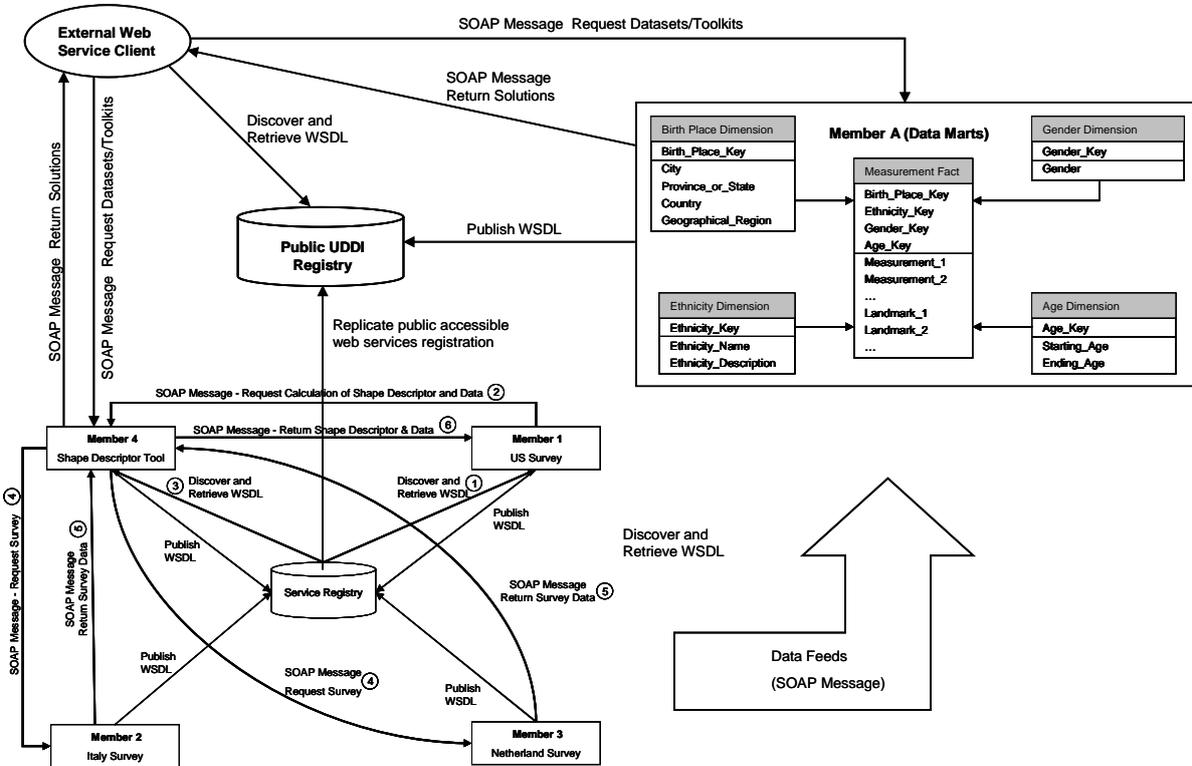


Figure 4. Hybrid web service / data mart model of WEAR

To implement this model, new capabilities offered by the specifications of the second generation of web services (WS-*) are needed in the areas of service coordination, single sign-on and role-based user authentication and authorization, transaction control for design process control, and so on. Even though there are still some uncertainties about WS-*, the general concept of this model is feasible. Regardless of how the analytical models will be developed and delivered, the approach of integration through XML web service and the concept of universal WEAR data feeders will make WEAR into not only a valuable anthropometric resource but also a building block to other applications.

6 CONCLUSION

This paper studies technical challenges faced by the WEAR integration. They can be categorized into three areas – integration of distributed and disparate data, universal representation of data, and autonomy for independent memberships. These challenges define the WEAR integration as a federated systems integration instead of a traditional enterprise or B2B. XML as a universal data format should serve as the foundation of the WEAR data representation. Several architecture options were reviewed, and XML web service was identified as the best solution due to its loosely coupling nature and service orientation. By addressing several perceived weaknesses in the implementation of the first-generation web service, the paper concludes that the widely supported first-generation XML web service is a workable solution to the short-term WEAR integration goal. This paper further proposed a hybrid web service/data mart model for the long-term WEAR objectives of offering analytical and design tools to the general public and private industries for solving engineering anthropometric problems. It also demonstrates that the web service integration of WEAR can serve as the data feeders to the analytical tools.

7 ACKNOWLEDGEMENT

The work is supported by the CARD Lab of the Bioscience and Protection Division, Human Effectiveness Directorate, United States Air Force Research Laboratory. The findings and conclusions in this report/presentation have not been formally disseminated by the Air Force and should not be construed to represent any agency determination or policy. This paper is approved for public release and distribution is unlimited. AFRL-WS 06-2697 (Nov 21 2006)

8 REFERENCES

Erl, T. (2006) *Service-Oriented Architecture – Concepts, Technology, and Design* (pp 291-292). Upper Saddle River, NJ: Prentice Hall

Microsoft Corporation (2006) *Configuring Linked Servers*, Retrieved August 20, 2006 from Microsoft MSDN Library, http://msdn.microsoft.com/library/default.asp?url=/library/en-us/adminsql/ad_1_server_4uuq.asp

Robinette, K., Blackwell, S., Daanen, H., Fleming, S., Boehmer, M., Brill, T., Hoeflerlin, D., and Burnsides, D. (2002) *Civilian American and European Surface Anthropometry Resource (CAESAR), Final Report, Volume I: Summary*, Technical Report AFRL-HE-WP-TR-2002-0169, United States Air Force Research Laboratory, Human Effectiveness Directorate, Bioscience and Protection Division, 2800 Q Street, Wright-Patterson AFB OH 45433-7947.