National Intelligence and the Integration Gap:
Implementing a Universal Data Model Across the Nation

By Len Silverston

Many IT professionals and business representatives have a difficult time envisioning how an enterprise data model can be implemented to integrate data across the applications within a single organization. Now, consider the need for a common national data model to integrate data across many major government agencies and commercial enterprises! The January 2002 article in this column (http://www.dmreview.com/master.cfm?NavID=55&EdID=4571) provided a high-level, common national data model, which was based upon **The Data Model Resource Book** (Wiley, 2001) series and customized toward national integration needs. That article also discussed issues involved in maintaining more integrated information, for example, what barriers exist in gathering national profiles of individuals, privacy issues, cooperation issues between government agencies, data matching/ accuracy issues and technology issues. While the implementation of such a model would be immensely valuable toward achieving many national goals such as counter-terrorism, there are many challenges and possible solutions regarding implementation of such a model. This article will address some of these issues and provide possibilities toward implementing a national data model for the purpose of integrating national information.

Before addressing more technical considerations, we need to understand that the greatest challenge in integrating information is getting people and organizations to share information. By giving up information, there is a perception that power is being given away. I assert that in order to gain real power (as a nation and as a world), it is necessary to give up power and be willing to share it. Only then, can we provide more accurate, consistent and complete view of the people, organizations, relationships and activities within our nation. With this integrated information, we could have tremendous capabilities towards better counter- terrorist intelligence, insight on preventing syndicated criminal activities and better information to make better national decisions that can positively affect our future as a nation.

Assuming that our government and nation is able, willing and committed to move toward this collaborative paradigm, how can we actually implement an integrated national information solution?

Data Integration Issues

Some of the issues that exist in integrating data from various government agencies and commercial enterprises are:

- Data exists within many formats, tools and platforms. For example, data exists in various types of databases, spreadsheets, communication logs, word processing documents, paper documents and electronic formats.
- Government and commercial organization terms and meaning of the data elements across government agencies may be quite different. For example, one government agency may use the term "suspect" to reference criminally suspicious individuals while another government agency may use the same term within a CRM context to refer to people that could be possible donations candidates to aid in helping terrorist victims (as in suspect, prospect and party that has donated).
- There are many inconsistencies in the data values for the same type of information between government enterprises. For example, an organization may have one corporate address maintained in one database and an inconsistent address for the same legal entity in a database maintained by another agency.
- Even within particular government agencies, there are usually many silos of information that maintain inconsistent and incomplete data.
- If there are inconsistencies between data values between and within various government and commercial organizations, there are no agreed-upon procedures or methods to identify, correct and/or resolve these differences in order to get more accurate and complete data.
- There are very few data stewardship programs in place across where specific people are assigned to oversee the quality of data at a national level and to facilitate synchronization, reconciliation and integration of data from various government and commercial organizations.
- There are a great number of different data architectures, operational data stores, data warehouse solutions, data marts and mechanisms for maintaining data across various government enterprises. This means that there could be a great variety of ways that organizations are maintaining data, for instance, maintaining the data in star schemas versus third normal form structures, and this could make it harder to integrate data.

Integrating data

With these issues in mind, the following section outlines some steps that are required to integrate data across various government and commercial organizations.
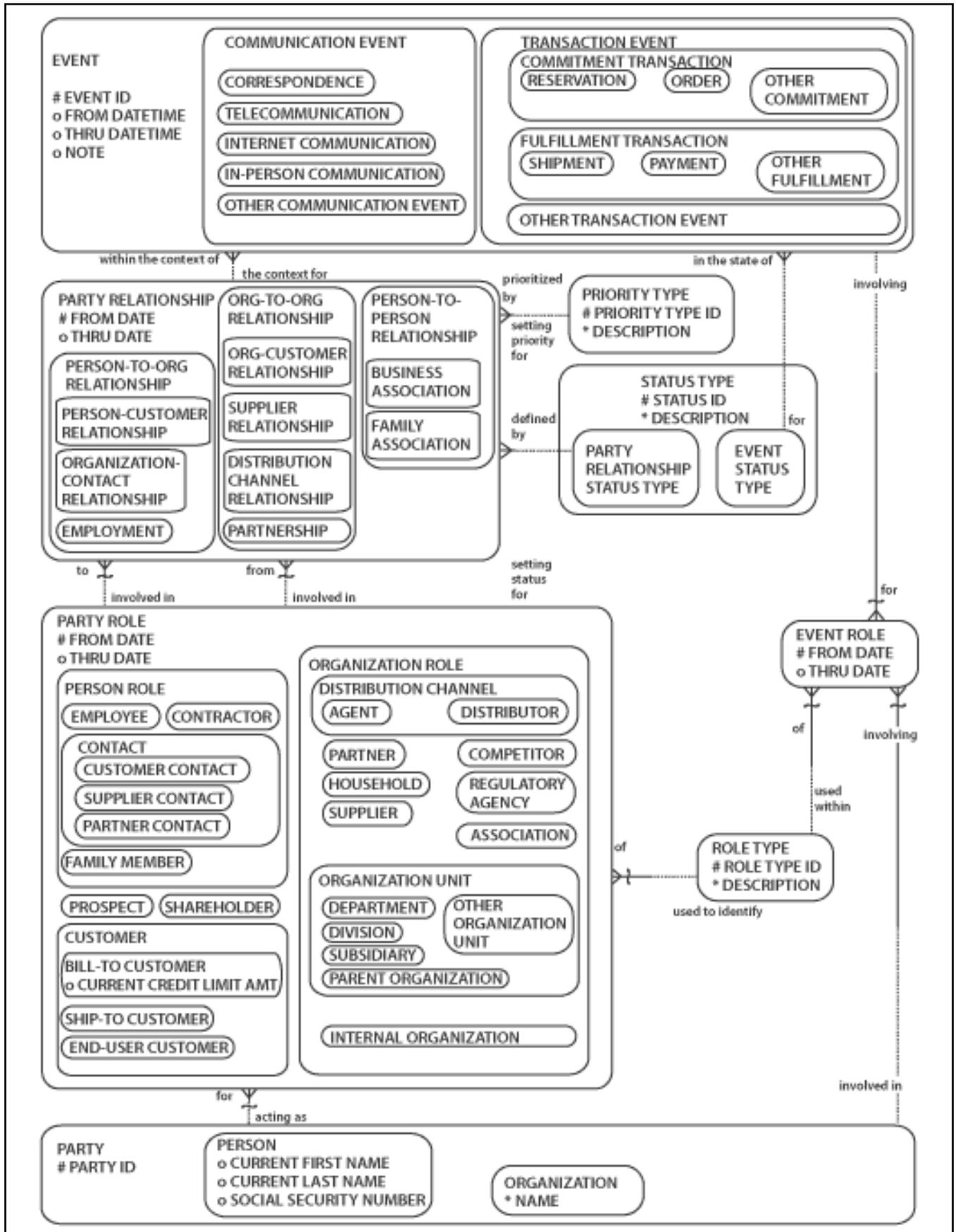
EVENT

EVENT ID
o FROM DATETIME
o THRU DATETIME
o NOTE

COMMUNICATION EVENT
- CORRESPONDENCE
- TELECOMMUNICATION
- INTERNET COMMUNICATION
- IN-PERSON COMMUNICATION
- OTHER COMMUNICATION EVENT

TRANSACTION EVENT

COMMITMENT TRANSACTION
- RESERVATION
- ORDER
- OTHER COMMITMENT

FULFILLMENT TRANSACTION
- SHIPMENT
- PAYMENT
- OTHER FULFILLMENT

OTHER TRANSACTION EVENT

within the context of / the context for

in the state of

involving

prioritized by

settling priority for

PARTY RELATIONSHIP
FROM DATE
o THRU DATE

PERSON-TO-ORG RELATIONSHIP
- PERSON-CUSTOMER RELATIONSHIP
- ORGANIZATION-CONTACT RELATIONSHIP
- EMPLOYMENT

ORG-TO-ORG RELATIONSHIP
- ORG-CUSTOMER RELATIONSHIP
- SUPPLIER RELATIONSHIP
- DISTRIBUTION CHANNEL RELATIONSHIP
- PARTNERSHIP

PERSON-TO-PERSON RELATIONSHIP
- BUSINESS ASSOCIATION
- FAMILY ASSOCIATION

defined by

PRIORITY TYPE
PRIORITY TYPE ID
* DESCRIPTION

STATUS TYPE
STATUS ID
* DESCRIPTION
- PARTY RELATIONSHIP STATUS TYPE
- EVENT STATUS TYPE

for

to / involved in

from / involved in

setting status for

PARTY ROLE
FROM DATE
o THRU DATE

PERSON ROLE
- EMPLOYEE
- CONTRACTOR
- CONTACT
  - CUSTOMER CONTACT
  - SUPPLIER CONTACT
  - PARTNER CONTACT
- FAMILY MEMBER
- PROSPECT
- SHAREHOLDER
- CUSTOMER
  - BILL-TO CUSTOMER
    o CURRENT CREDIT LIMIT AMT
  - SHIP-TO CUSTOMER
  - END-USER CUSTOMER

ORGANIZATION ROLE
- DISTRIBUTION CHANNEL
  - AGENT
  - DISTRIBUTOR
- PARTNER
- COMPETITOR
- HOUSEHOLD
- REGULATORY AGENCY
- SUPPLIER
- ASSOCIATION
- ORGANIZATION UNIT
  - DEPARTMENT
  - DIVISION
  - SUBSIDIARY
  - OTHER ORGANIZATION UNIT
  - PARENT ORGANIZATION
- INTERNAL ORGANIZATION

EVENT ROLE
FROM DATE
o THRU DATE

of / involving

used within

ROLE TYPE
ROLE TYPE ID
* DESCRIPTION

used to identify

of

involved in

for / acting as

PARTY
PARTY ID

PERSON
o CURRENT FIRST NAME
o CURRENT LAST NAME
o SOCIAL SECURITY NUMBER

ORGANIZATION
* NAME

Figure1: A Universal Data Model for National Profiling

Define data requirements for integration. There is a tremendous amount of national data that could be integrated and pooled into a national repository to better serve our nation. In order to make progress, it is critical to have an incremental approach that integrates data from various sources one piece at a time. The first step involves identifying data integration possibilities, prioritizing these possibilities based upon the impact that each of them provided and setting up a phased integration plan to implement data integration.

Agree upon a common data model standard. The next step would be to agree on common data structures that would define consistent terms and structures for maintaining common information. Examples of common information among agencies include personal demographics, organization firmographics, contact information, relationships between various people and organizations, activities that have occurred over time, licensing data, government applications and communications that have occurred between various parties. A possible common national data model was provided and explained in the last article for this column and is shown for easy reference purposes in Figure 1.

Map data from selected source systems into common data model. After various government and commercial organizations agree upon a common data structure they could then map their data from each of their source systems to this common structure in an effort to bring data together in a common format.

Figure 2 provides an illustration of data from several government agencies being integrated and synchronized using a physical implementation of the common data model presented in the last article. The NATIONAL ID TO APPLICATION CROSS REF relates each of the APPLICATION TABLEs within each APPLICATION to a PARTY entity in order to identify all the information for a party across applications.

For example, assume that a high priority was to have an integrated profile of suspicious individuals and to start by integrating immigrations and FBI information. If the Immigration Naturalization Service (INS) stored records of people that have immigrated to the U.S. in an IMMIGRANT table and the FBI maintained wanted individuals in a SUSPECT table, then they could both map their data to a PARTY table (or PERSON table) and maintain a record for each party (or person) along with their respective roles in the PARTY ROLE table.

There are two main methods for mapping data from a source system to a common, abstract data schema such as the one suggested in the common national data model:

- Put foreign keys in the source structures that relate to a more abstract data structure in the common national model and that provide a mechanism for bringing together data about the same thing. For instance, allow a party_id to be entered as a foreign key attribute within both the IMMIGRANT and SUSPECT data structures. There would need to be a consistent algorithm to generate the same party id for the same individual or organization existing in more than one data store.
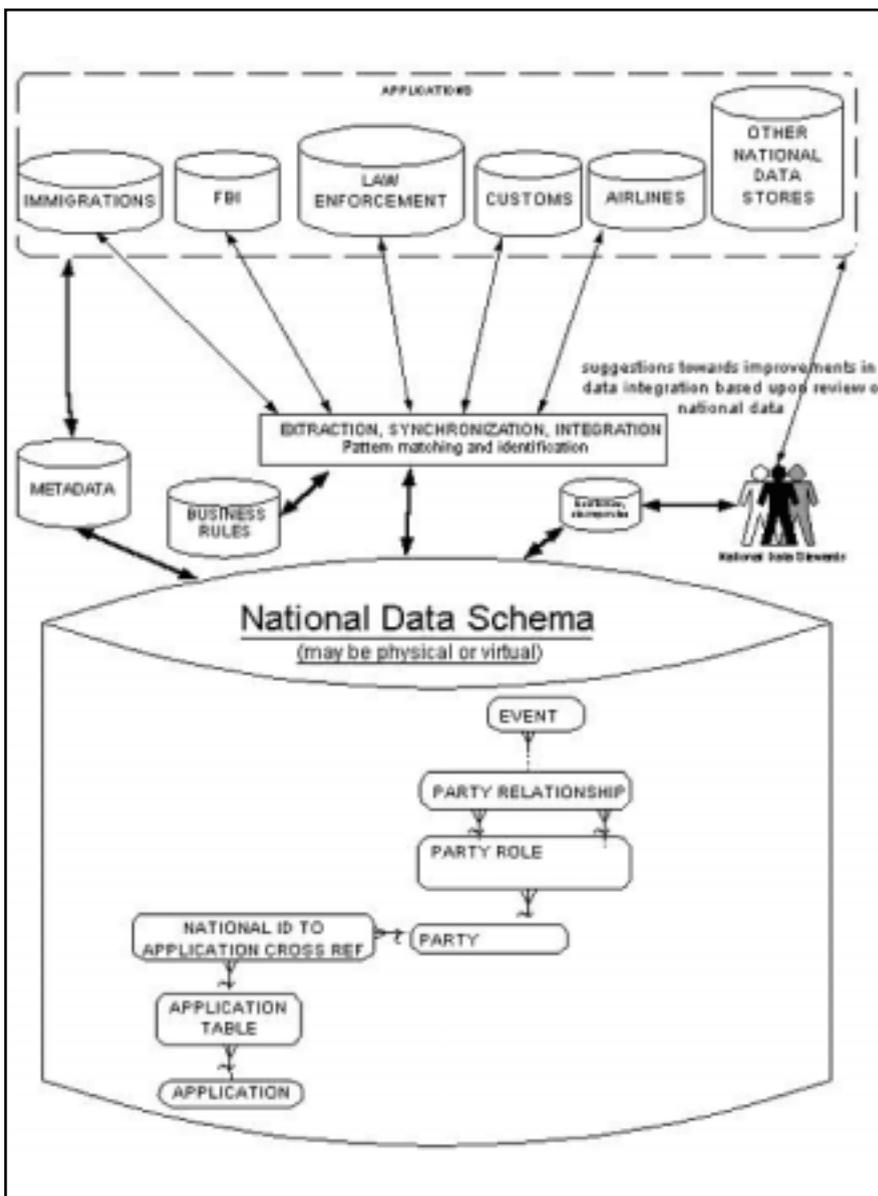

Figure 2: National Data Model Implementation

- Develop cross-reference tables that map the primary key values within source system tables to the abstract primary key in the common national model. This latter method is the one shown in the national data schema in Figure 2.

Identify the various inconsistencies that exist in the data. Once data is moved into a common format, more complete information may be obtained and inconsistencies can be identified. Using the previous example, the common data structure could provide a more complete profile of a person that was both an immigrant and an FBI suspect by revealing his or her immigration history as well as suspicious activities for which they have been involved. As illustrated in Figures 1 and 2, the common model could maintain a single PARTY of subtype PERSON that had two PARTY ROLES of IMMIGRANT and SUSPECT and maintain a consolidated history of EVENTs showing the type of involvement the party played in the event (the EVENT ROLE).

Define Algorithms for Combining Records

How would the integrated data schema know when to combine records? For instance, if it were decided to integrate party records from many sources, how would the system identify records from multiple systems which reference the same person? The immigration system may have a record for "Joe Smith" at 10037 Niven Street in Baltimore, MD, and the FBI system may have a record for "Joe Smithe" at 10937 Niven Street in Baltimore, MD. Is this the same person?

Business rules are needed to identify when to combine records into the common national data schema, when to maintain these records as separate instances of the common entities and when to flag certain records as possible matches. Thus the architecture in Figure 2 has a process that performs pattern matching using these national business rules that identify uniqueness are needed. This would allow, as part of the system, synchronization and integration of national information from several government systems.

The example for Joe Smith would most likely be flagged in a discrepancies database so that data stewards were aware of possible inconsistencies and could do research to resolve possible errors.

Identify Procedure to Handle Inconsistencies

What if the information for the same person is different? When combining information from several sources, there are bound to be "if" inconsistencies in the data. If the records for Joe Smith were combined, there were still inconsistent addresses of 10037 Niven Street in Baltimore, MD, versus 10937 Niven Street in Baltimore, MD. Which one would be maintained in the common government system?

There are a number of ways to handle this situation. If the business rules identified that this was the same person with different inconsistent addresses, then the common data store could just maintain the party_id key for John Smith and cross reference it to each of the application key values stored in each of the source systems. This would allow the national data schema to bring up consolidated information on the party, John Smith, and show the many variations of data (which may be inconsistent) that are related to this party.

Another alternative is to maintain structures in the common data store to capture many variations of the same data. In our example, the common national data schema could allow for any number of additional addresses and identify its source. This common schema would also need to allow multiple values for other types of inconsistent data such as name inconsistencies, demographic inconsistencies, contact information inconsistencies and so on.

Design Feedback Loop

Many data integration specialists contend that data should be cleansed before being passed into the common integrated data store. For example, if it was determined that Joe Smith's name was misspelled in the INS database (since the FBI record was deter-

mined to be the system of record for name information), then the cleansed data maintained in the common national data schema would be the name from the FBI record.

A more powerful paradigm would be to identify data inconsistencies and to provide data stewards with tools they could use to help cleanse data inconsistencies at the root level, namely at each of the source systems that would feed the common national schema. Notice the data discrepancies data store in Figure 2 and the feedback loop from the data stewards back to the operational systems.

Design Technical Architecture

There are a number of alternatives and technical design considerations regarding integrating national data. For instance, the common national data schema could be either a virtual schema or a physical data store that maintains a consolidated archive of data from the various constituent systems.

The method for passing information back and forth from each of the source systems could be XML, and the data could be passed using common XML schemas that correspond closely to the common national data model structure. This would promote standard structures for passing data and consistent meaning regarding the data being passed.

Business rules need to be captured and integrated into the technical architecture. For instance, there should be a complete set of business rules identifying when to combine the records of people when they are in multiple sources.

There are many middleware solutions that serve as excellent agents for passing data back and forth through their messaging queuing systems in order to synchronize and integrate data, for example MQ Series or Tibco. Thus, it would be wise to standardize on a middleware solution enabling a consistent architecture to integrate national information.

Implement Solution in Iterations

Thus, our government could start integrating information from several

agencies and commercial enterprises. Since the number of government agencies and commercial enterprise is huge, the only sensible approach is to start with the highest priority agencies and incrementally add additional organizations' information as time goes on. Perhaps a first phase of integration could be integrating two agencies such as Immigration and FBI, then customs information, then NSA information, then airlines information and so on. A small scope of key information could be specified for each organization and additional data could be brought in for each organization in future stages. The net result would be a more and more powerful, integrated source for national information and intelligence.

Implementation of the common national model is technically feasible. While there are many considerations for integrating national information into a common data structure, there are numerous solutions that provide an integrated view of national information. The implementation of a common national model could make a significant difference in providing much better intelligence towards securing our nation, combating terrorism and improving national decisions.

*Len Silverston is a data management consultant with more than 20 years of experience in helping enterprises integrate data. He is the author of the best-selling **The Data Model Resource** Book series (Wiley, 2001), which describes more than 230 integrated, reusable generic and industry-specific data models. Silverston has developed extensive software versions of these data models, some of which are now licensed worldwide by Microsoft and others that are available for licensing directly. Silverston's company, Universal Data Models, provides consulting, training and software to jump-start data modeling and data warehouse design efforts while increasing design quality and facilitating data integration. Silverston can be reached at lsilverston@univdata.com.*