Designing Data Collection, Processing, and Dissemination Instruments With Reusable Metadata for the U.S. Census Bureau's 2002 Economic Census Initiative, Using XML and Web Services

Steven A. Schafer and Roy S. Rogers, IV
Fenestra Technologies Corporation, 20410 Century Blvd. Ste. 230, Germantown, MD 20874

Abstract

Fenestra Technologies Corporation recently completed a three-year project in conjunction with the United States Census Bureau. The goal of the project was to create software for semi-automated generation of both paper and electronic survey forms for use in the 2002 Economic Census. The result was a multi-component software package, the Generalized Instrument Design System (GIDS). GIDS is a metadata-driven and XML-based tool for the efficient creation of survey forms. This paper focuses on the XML-related aspects of GIDS, relates our positive and negative experiences in that regard, and discusses the future direction that we have planned for GIDS as it evolves into a more broadly-scoped information management tool.

I. Introduction

Every five years, the Economic Directorate of the United States Census Bureau, a division of the U.S. Department of Commerce, conducts a wide-ranging census of economic activity that involves over 650 individual survey forms averaging 10–12 pages each (see http://www.census.gov/epcd/www/econ2002.html). These surveys are distributed to over 5 million U.S. businesses. The survey forms measure current economic activity, and so the specific questions vary from one census cycle to the next to reflect the changes in the U.S. economy. Consequently, the forms must be redesigned for each census.

Creating and managing these survey forms is a significant undertaking, involving the combined efforts of hundreds of people over a two- to three-year timeframe. Historically, the survey forms have been created individually by hand. Obviously, this process is labor-intensive, error-prone, and difficult to translate to the electronic world of online surveys. The subject matter experts who create the survey questions and design the forms mark up paper drafts of the form layouts and send them to a separate department where graphic artists use conventional graphics drawing software to compose the forms. The turnaround time for a single iteration of the layout/edit cycle is anywhere from several days to several weeks.

For the 2002 Economic Census, the Census Bureau contracted with Fenestra to design and implement solutions to two separate problems. First, it wanted to streamline the process of survey design and production so that subject matter experts would have essentially real-time feedback as they design and edit their form layouts. Second, it wanted to drive the creation of both paper and electronic (online) surveys from a single, common data repository containing the content and layout information for the survey forms.

The end result of Fenestra's work with the Census Bureau is the Generalized Instrument Design System (GIDS). A block diagram of GIDS is shown in Figure 1.

GIDS consists of several modules:

- Forms Designer – The Forms Designer is used by the subject matter experts to lay out "custom-formatted" sections of forms; these are sections that are not amenable to fully automated layout.

- Autoformatter – The Autoformatter automatically lays out regular, repeating sections of forms, based on a set of layout rules and templates.

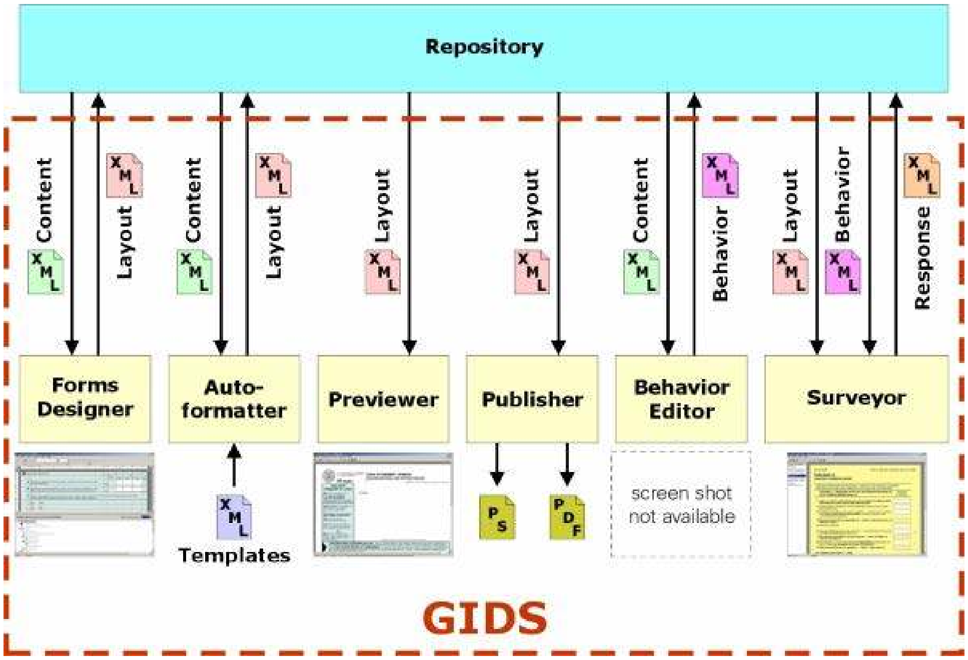


Figure 1. A block diagram of GIDS.

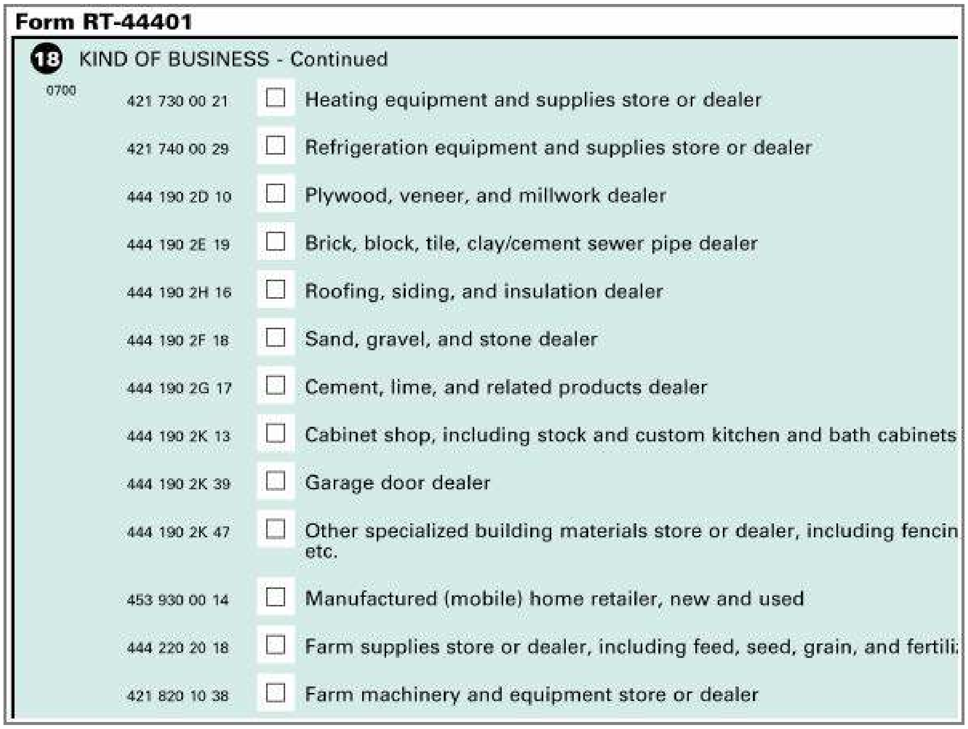
- Previewer – The Previewer provides visual feedback to the designers so that they may inspect forms in their entirety, with custom-formatted and autoformatted sections combined along with boilerplate into the final forms.

- Publisher – The Publisher takes form layouts destined for paper output and produces PostScript® or Adobe® Portable Document Format (PDF) files for printing, either with the Census Bureau's in-house printing facilities (for limited-run forms), or by commercial printers (for large-quantity production).

- Behavior Editor – Electronic forms have behavior (auto-calculated response fields, navigation, data validation) in addition to visual layout. The Behavior Editor is the means by which the subject matter experts attach behaviors to the various items on a form.

- Surveyor – The Surveyor is the electronic equivalent of the Publisher—it presents an electronic form to a respondent, collects the respondent's response data, and securely transmits those data back to the Census Bureau.

The next three figures show examples of rendered output from GIDS; all three examples are of output destined for paper, rather than online, forms. Figure 2 shows a portion of a simple autoformatted (i.e., rule-based) layout. Figure 3 shows a more complex hierarchical autoformatted layout. This example also illustrates the auto-numbering and item cross-referencing features built into the GIDS software. The various "Continued" headers are also generated automatically, as required, at page breaks. Figure 4 shows a portion of a custom-formatted section, one that was laid out "by hand" using the Forms Designer. This example also shows some of the non-text capabilities of GIDS, such as embedded graphics (the barcode at the left edge of the form is generated programmatically at form assembly time, and embedded as a bitmap image). On a page-count basis, approximately 80% of the Economic Census survey forms consist of autoformatted sections; the remainder are custom-formatted.



Form RT-44401

18 KIND OF BUSINESS - Continued

0700	421 730 00 21	□	Heating equipment and supplies store or dealer
	421 740 00 29	□	Refrigeration equipment and supplies store or dealer
	444 190 2D 10	□	Plywood, veneer, and millwork dealer
	444 190 2E 19	□	Brick, block, tile, clay/cement sewer pipe dealer
	444 190 2H 16	□	Roofing, siding, and insulation dealer
	444 190 2F 18	□	Sand, gravel, and stone dealer
	444 190 2G 17	□	Cement, lime, and related products dealer
	444 190 2K 13	□	Cabinet shop, including stock and custom kitchen and bath cabinets
	444 190 2K 39	□	Garage door dealer
	444 190 2K 47	□	Other specialized building materials store or dealer, including fencin etc.
	453 930 00 14	□	Manufactured (mobile) home retailer, new and used
	444 220 20 18	□	Farm supplies store or dealer, including feed, seed, grain, and fertili
	421 820 10 38	□	Farm machinery and equipment store or dealer

Figure 2. A portion of an autoformatted section of a survey form.

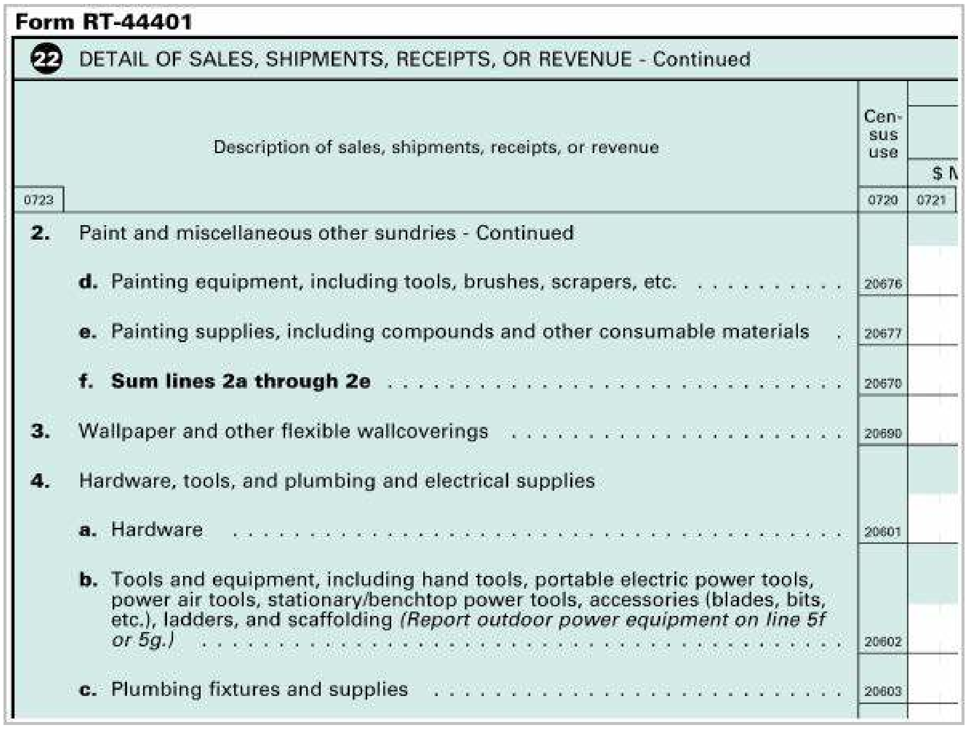
Figure 5 shows a portion of an electronic survey, as displayed in the Surveyor application. As can be seen, the general "look and feel" of the electronic form resembles that of the paper version, but there are numerous small differences, the result of tailoring the form display to the characteristics of the presentation medium.

II. Rationale for using XML-based technologies

When we first began work on GIDS approximately four years ago, we made the decision to use XML as the format for information storage and interchange with other systems. Our decision was based on several factors. First, XML is "human-readable." That is, a person can look at an XML file and deduce the nature of its content, at least in principle. In practice, at least some information in an XML file—a graphical image, for example—has to be encoded in some sort of non-human-readable format for efficient processing. Even in that case, however, it is still possible to include documentary metadata alongside the non-human-readable information that explains how it is encoded.

Second, because XML is a well-defined open standard, it increases the feasibility and practicality of data interchange among disparate processing systems. In the case of GIDS, those external systems included relational databases, and legacy mainframe-based systems that continue to rely on the same fixed-length 160-column record formats that were first developed over thirty years ago.

Third, the World Wide Web consortium, which develops most XML-based standards, is a thriving nexus of innovation in information management and processing. As the repertoire of XML-based standards evolves, we begin to reach a point where constructing a fully-customized, full-fledged data processing tool becomes largely a matter of choosing from a variety of existing XML-based components and plugging them together.

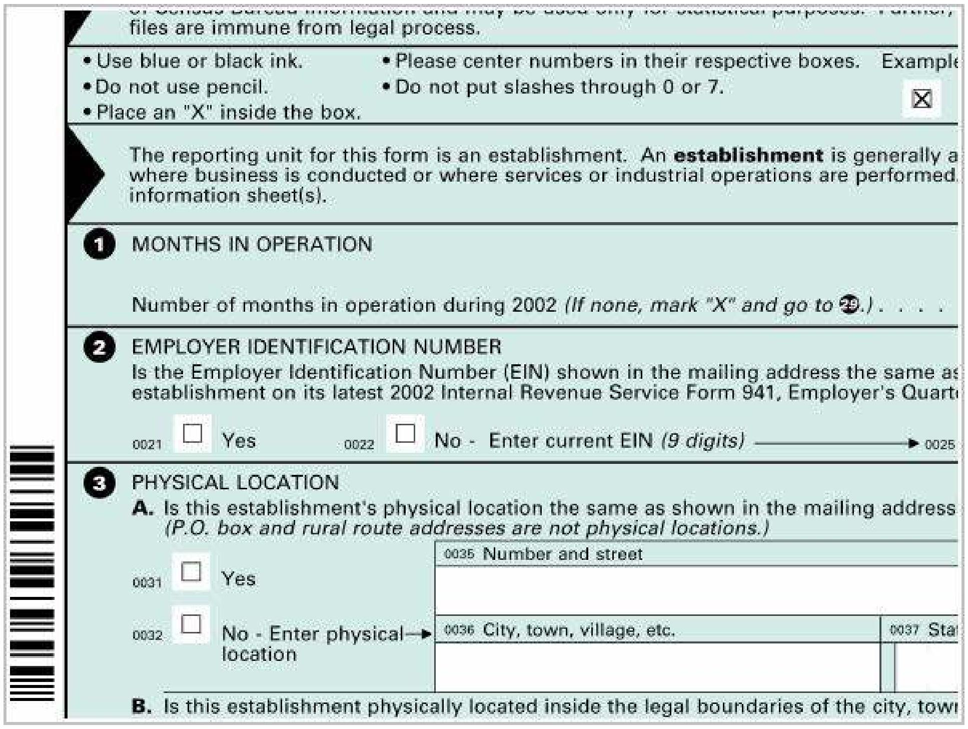


Form RT-44401

22 DETAIL OF SALES, SHIPMENTS, RECEIPTS, OR REVENUE - Continued

0723 Description of sales, shipments, receipts, or revenue	0720 Census use	0721 $ N
2. Paint and miscellaneous other sundries - Continued		
d. Painting equipment, including tools, brushes, scrapers, etc.	20676	
e. Painting supplies, including compounds and other consumable materials .	20677	
f. Sum lines 2a through 2e	20670	
3. Wallpaper and other flexible wallcoverings	20690	
4. Hardware, tools, and plumbing and electrical supplies		
a. Hardware	20601	
b. Tools and equipment, including hand tools, portable electric power tools, power air tools, stationary/benchtop power tools, accessories (blades, bits, etc.), ladders, and scaffolding *(Report outdoor power equipment on line 5f or 5g.)*	20602	
c. Plumbing fixtures and supplies	20603	

Figure 3. A portion of a hierarchical autoformatted section of a survey form.



of Census Bureau information and may be used only for statistical purposes. Further, files are immune from legal process.

- Use blue or black ink.
- Do not use pencil.
- Place an "X" inside the box.
- Please center numbers in their respective boxes.
- Do not put slashes through 0 or 7.

Example ☒

The reporting unit for this form is an establishment. An **establishment** is generally a where business is conducted or where services or industrial operations are performed. information sheet(s).

1 MONTHS IN OPERATION

Number of months in operation during 2002 *(If none, mark "X" and go to 29.)* . . . .

2 EMPLOYER IDENTIFICATION NUMBER

Is the Employer Identification Number (EIN) shown in the mailing address the same as establishment on its latest 2002 Internal Revenue Service Form 941, Employer's Quart

0021 □ Yes 0022 □ No - Enter current EIN *(9 digits)* → 0025

3 PHYSICAL LOCATION

A. Is this establishment's physical location the same as shown in the mailing address *(P.O. box and rural route addresses are not physical locations.)*

0031 □ Yes

0032 □ No - Enter physical location →

0035 Number and street

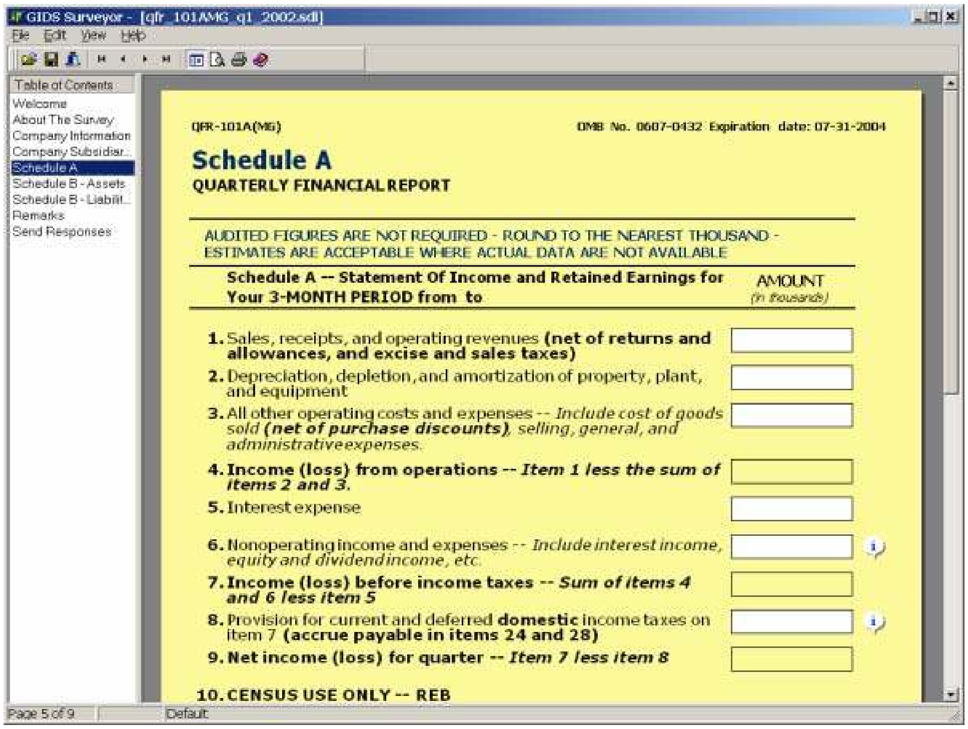
0036 City, town, village, etc.

0037 Sta

B. Is this establishment physically located inside the legal boundaries of the city, tow

Figure 4. A portion of a custom-formatted section of a survey form.

Interestingly, when we first proposed the use of XML in GIDS (in early 2000), we were met with strong opposition from some of the other parties involved in the project. This opposition could be traced to lack of familiarity along with lack of XML support in those parties' favorite software construction tools. But by the time it was necessary to connect the various software components together (early 2002), XML support had become widespread among commercial software vendors, and XML-based connectivity between GIDS components was accomplished without difficulty.



QFR-101A(MG) OMB No. 0607-0432 Expiration date: 07-31-2004

Schedule A

QUARTERLY FINANCIAL REPORT

AUDITED FIGURES ARE NOT REQUIRED - ROUND TO THE NEAREST THOUSAND - ESTIMATES ARE ACCEPTABLE WHERE ACTUAL DATA ARE NOT AVAILABLE

Schedule A -- Statement Of Income and Retained Earnings for Your 3-MONTH PERIOD from to	AMOUNT *(in thousands)*
1. Sales, receipts, and operating revenues **(net of returns and allowances, and excise and sales taxes)**	
2. Depreciation, depletion, and amortization of property, plant, and equipment	
3. All other operating costs and expenses -- *Include cost of goods sold **(net of purchase discounts)**, selling, general, and administrative expenses.*	
4. Income (loss) from operations -- *Item 1 less the sum of items 2 and 3.*	
5. Interest expense	
6. Nonoperating income and expenses -- *Include interest income, equity and dividend income, etc.*	
7. Income (loss) before income taxes -- *Sum of items 4 and 6 less item 5*	
8. Provision for current and deferred **domestic** income taxes on item 7 **(accrue payable in items 24 and 28)**	
9. Net income (loss) for quarter -- *Item 7 less item 8*	
10. CENSUS USE ONLY -- REB	

Figure 5. An electronic survey.

At the time we chose XML, we also selected eXtensible Stylesheet Language Formatting Objects (XSL-FO) as the means of describing the text and graphics on the survey forms. However, it became apparent fairly early on that the XSL-FO specification would not be sufficiently complete by the time we needed to use it. For that reason, we developed our own interim formatting language, SFO, which we based loosely on the existing XSL-FO specification, with the intention being that we would eventually merge SFO with XSL-FO.

A key requirement of the Economic Census is precise control over all aspects of the visual appearance of the forms: typography, position of elements, colors, etc. SFO was designed with this requirement in mind, and many of the design decisions derive directly from it. All of the example forms shown in Figures 2–5 are stored in SFO format and then rendered on paper or on screen as needed.

III. Rationale for using web services

Most interaction and data interchange on the World Wide Web involves at least one active human participant (the person sitting in front of a computer, using a web browser). The web infrastructure, however, is also amenable to computer-computer data interchange, without human intervention, and this is the motivation for the creation of web services.

GIDS uses web services technology in a fairly simple and straightforward manner: The Surveyor application, running on a respondent's computer, automatically connects to a Census Bureau server over the Internet to retrieve the appropriate list of forms (based on the respondent's identification code), the forms themselves, and any prelisted survey data, such as name and address. Similarly, the Surveyor uses the same web services framework to upload response data back to the Census Bureau server.

The XML-based standards used for these transactions are the Secure Sockets Layer (SSL), which provides for encryption and data security, and the Simple Object Access Protocol (SOAP), which describes the actual mechanics of data interchange. Although the use of SOAP, and web services in general, was limited, the intent is to broaden the scope in the future. For example, it may be possible in the future for a respondent's accounting application to communicate directly with a Census Bureau server to provide financial information for a survey, while at the same time the respondent's human resources management application is providing personnel information for the same survey.

IV. The role of metadata

Although we describe GIDS as being metadata-driven, the kinds of metadata used by GIDS are generally simple (field types and lengths, question captions, etc.). In order to extend and generalize the concepts used in the design of GIDS across the overwhelming diversity of business and government applications, the availability of high-quality metadata is an absolute requirement. This is especially important in the context of increasingly complex computer-computer data interchange via web services; The only thing worse than having to manually enter survey data for a thousand separate business establishments is having your computer do it for you—but send all of the wrong data. The development of high-quality metadata is probably the single largest challenge facing the information and knowledge management communities, for today and for the foreseeable future.

V. Current work

As mentioned previously, we had designed the SFO layout and formatting language with the intention of merging it into XSL-FO when the latter standard became fully stabilized. But as we accumulated experience with GIDS and SFO, it became clear that XSL-FO was probably not the best choice for our purposes. Instead, we have decided to use the Scalable Vector Graphics (SVG) standard as the basis for layout and formatting in the next generation of GIDS. We will also use XForms to represent data sets and to provide the use interface controls on electronic forms. Both SVG and XForms rely on other World Wide Web standards, such as XPath and XPointer for navigation within an XML document, XLink for linking between documents, Cascading Style Sheets (CSS) for visual formatting, etc. We will increasingly apply SOAP and web services in more demanding inter-computer data interchange roles.

By basing GIDS on a number of well-defined and open standards, our goal is to design a system that is freely accessible to third-party software components, including those written by end users (e.g., survey respondents) as well as by other software vendors.

Fenestra is also involved in the "eForms for eGov" initiative whose goal is to achieve widespread adoption of electronic data collection and dissemination throughout the Federal Government. One of the key pieces of such a system, and one that currently lacks an acceptable standard as a basis, is a comprehensive information and knowledge management system. Through eForms/eGov and other avenues, Fenestra is pursuing the creation of a standard for information management.

VI. Summary

Overall, the GIDS project and the 2002 Economic Survey stand as an extremely successful example of the use of XML-based technologies and the Internet . At the time of this writing, over NN paper surveys and NN electronic surveys have been submitted to the Census Bureau. The use of XML as the standard interchange format proved to be a wise decision, as the problems encountered when transferring data between disparate systems, often the most difficult part of a data collection and dissemination project, were minimal. The lack of availability of high-quality metadata represents the most significant hurdle that must be overcome before the goal of broad cross-agency and cross-business data interchange can be achieved.