The NIST Data Gateway: Providing Easy Access to NIST Data Resources

Dorothy M. Blakeslee[1], Angela Y. Lee[2], and Alec J. Belsky[3]

National Institute of Standards and Technology, Standard Reference Data Group, 100 Bureau Drive, Stop 2310, Gaithersburg, MD 20899-2310, USA

[1]*Email: dorothy.blakeslee@nist.gov*
[2]*Email: angela.lee@nist.gov*
[3]*Email: alec.belsky@nist.gov*

Abstract

The National Institute of Standards and Technology (NIST) maintains a wide range of scientific and technical data resources, including free online data systems and PC databases available for purchase. However, many people are not familiar with these various NIST data collections and the types of data they contain. To help scientists, engineers, and the general public find out quickly and easily whether data they need are available at NIST, NIST has built a web portal to NIST data resources. The first version of this portal, the NIST Data Gateway (http://srdata.nist.gov/gateway), provides easy access to 26 online NIST data systems and information on 48 NIST PC databases. NIST Data Gateway users can specify a keyword, property, or substance name to find the NIST data resources that contain standard reference data meeting their search criteria. When users find a data resource they want to use, links are provided so they can access or order that resource. In this paper, we describe how Version 1.0 of the NIST Data Gateway was built and discuss some of the issues that arose during the design and implementation stages. We include experience we gained that we hope will be useful to others building data portals. We also discuss future plans for the NIST Data Gateway, including efforts to provide access to additional NIST data resources.

Keywords: Data, Database, Interface, Online system, Portal, Scientific data, Standard reference data, Technical data, Web

1. Introduction

The last ten or fifteen years have seen tremendous growth in the amount of scientific and technical data that is available electronically. At the National Institute of Standards and Technology (NIST) alone, there are over 100 PC and online databases, covering a broad range of data in a variety of disciplines. To improve the accessibility of the data in these various databases, NIST has initiated a project to develop and implement a web portal to NIST data. The goal is to serve both expert and non-expert users by helping them find out quickly and easily whether NIST has the data they are seeking.

This project is being carried out in phases, each succeeding phase building on the work of the previous. The first phase, which began in 2000, concentrated on the design and development of the initial concept for the data portal. The NIST Standard Reference Data Group (SRDG) investigated various options and strategies. After completing their analysis, the project members decided to develop a prototype portal, the NIST Data Gateway (referred to hereafter as the "Gateway"), to test the viability of their initial design for the system, including the web interface, search options, and implementation methods. This prototype, which was completed in September 2000, was based on a content management database populated with indexing information from a sample set of NIST databases.

Following the successful release of the prototype, the second phase of the project was undertaken. A Gateway Team was created in NIST SRDG to evaluate the prototype and plan the next steps. The Gateway Team was tasked with using the knowledge and experience gained while developing the prototype, as well as feedback from reviewers of the prototype, to improve upon the original design. The result was a redesigned system, the NIST Data Gateway Version 1.0, which expanded the number of NIST online and PC databases accessible via the Gateway. Version 1.0, which underwent extensive testing and beta review before it was officially released to the public in November 2001, can be accessed at http://srdata.nist.gov/gateway. Alternatively, users can go to the NIST Home Page (http://www.nist.gov), click on the Databases link to bring up the NIST Scientific Databases page, and then click on the Data Gateway link to access the Gateway.

In addition to providing links to 26 free online NIST databases, Version 1.0 of the Gateway provides access to information on 26 PC-based standard reference databases and 22 special databases that can be ordered from the NIST Standard Reference Data Group. For experts or experienced users, who are familiar with NIST databases and know the name of the online or PC database they are interested in, the Gateway provides a link either to the online database or to information on the PC database. For non-experts or inexperienced users, an easy-to-use search system featuring keyword, property, and substance name search options is provided so users can easily find out whether any of the NIST databases indexed in the Gateway contain the type of data they are seeking. After the search is completed, links are provided for any databases matching the search criteria.

Future plans for the Gateway include both enhancing the functionality provided and increasing the number of databases indexed. The goal is to add more NIST databases incrementally, as they are indexed, and to expand the scope of the Gateway to other data resources, such as computational software, so users will have easy access to as many NIST data resources as possible.

2. Building the NIST Data Gateway

2.1 Content Management Database

One of the most important decisions that needed to be made before the Gateway could be built was determining how to implement the portal searches. Because of the nature of the content of the NIST data resources, it was decided that a search engine with functionality

based on general text searching of web pages would not be reliable enough and that search strategies based on attributes specifically chosen to be appropriate to scientific and technical data would be more reliable, i.e., less likely to return misleading results. Therefore, a system was built to provide search options based on predetermined keywords, properties, and substance names.

This system is implemented through a content management database that contains metadata on all the databases indexed by the Gateway. The content management database was designed and built by members of the Gateway Team so that it could be tailored to reflect accurately the contents of the NIST databases. Metadata stored for each database include the database name, version number, brief description, type (PC or Web; free or for purchase), scientific contact, and Uniform Resource Locator (URL). For online databases, the URL is the address of the opening page of the database. For PC databases, the URL is the address of the web page in the NIST Scientific and Technical Databases Web Site (http://www.nist.gov/srd) that describes the PC database and provides a link to a web page where it can be ordered.

Indexing terms are stored in the content management database for all the databases available in the Gateway. All the databases in Version 1.0 are indexed by keyword. Many are also indexed by property and substance name. The developers of the online databases provided the indexing terms for their databases. For example, approximately 35,000 substance names and approximately 90,000 synonyms came from the NIST Chemistry WebBook (http://webbook.nist.gov/chemistry) and approximately 14,000 substance names came from the Protein Data Bank (http://nist.rcsb.org/pdb). Indexing terms for the PC databases were largely derived from those provided in the NIST Scientific and Technical Databases Web Site.

After the initial set of indexing information was collected, the database names and descriptions were checked for accuracy and completeness. The keywords and properties were sorted and printed in two formats, one ordered by index term and one ordered by database, then reviewed for accuracy and data consistency. For example, if one database with spectral data was indexed by the term *spectra*, checks were made to ensure that all databases with spectral data were also indexed by *spectra*. In addition, the properties were classified into major and minor categories so that a property search by category could be implemented. For example, the property *electron affinity* was assigned to the major category *Atomic, Molecular, and Optical Physics* and to the minor category *atomic property*.

The NIST Chemistry WebBook was a resource for providing synonyms for substance names. The substances in each database were compared programmatically with the substances in Chemistry WebBook, using either the Chemical Abstracts Service (CAS) Registry Number or the substance name as a key. Any substances that matched those in the Chemistry WebBook were assigned the same synonyms as indexing terms for the Gateway. In addition, some validity checks were made on the substance names, but it was not possible to check each individual name manually.

Next the Gateway Team members consulted with the authors of all the NIST databases included to verify that their databases were indexed correctly by keyword and property. After the review was completed, the indexing terms were finalized in the content management database. A total of 372 keywords and 451 properties as well as over 140,000 substance names and synonyms are included in Version 1.0.

2.2 Web Interface

The pages in the Gateway web interface are designed to follow NIST conventions for the overall layout and style. There is a banner identifying the Gateway across the top of each page. Directly under the banner are links to Gateway Home (Welcome page), three search options, and help information. The left column of each page contains a list of links to NIST data resources, information about Standard Reference Data, feedback/comments page, and site information. The contents of the banner and left column remain fixed for each page in the Gateway site. The main portion of the page varies according to what option the user chooses. See Figure 1 for a display of the opening screen.

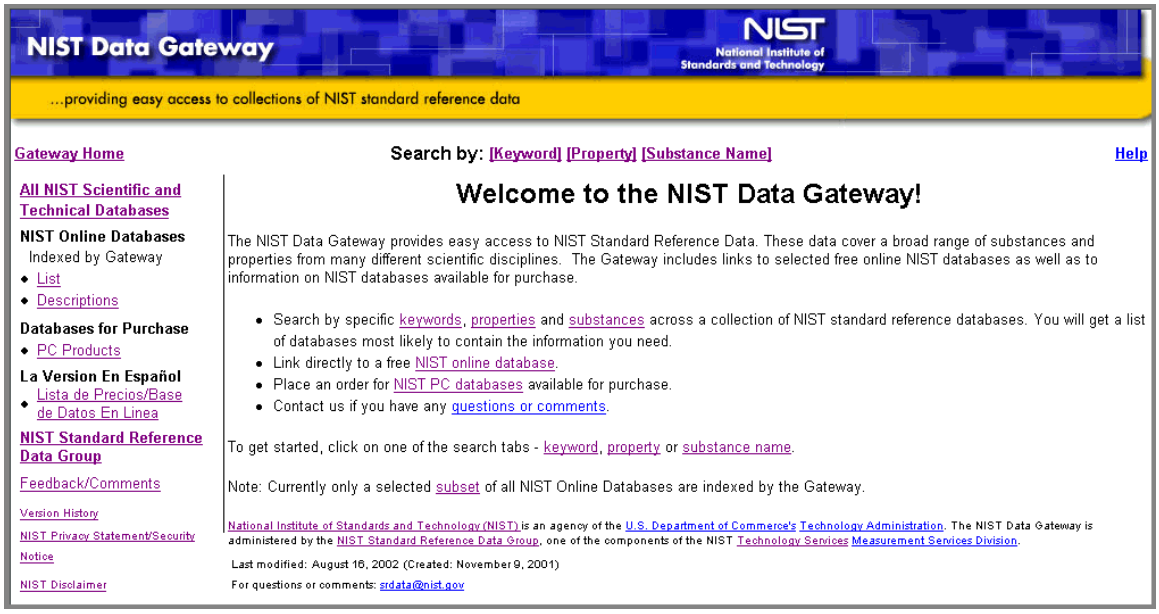

Figure 1. Opening Screen of the NIST Data Gateway

Many of the links in the left column are intended primarily to help experienced users, i.e., users who are already knowledgeable about many of the NIST data resources. For example, a user may already be familiar with the CODATA Fundamental Physical Constants site (http://physics.nist.gov/PhysRefData/contents.html), but not be able to remember its URL. In this case, the user could click on the List link under NIST Online Databases in the left column to bring up an alphabetical list of all online databases included in the Gateway, and then click on the link for the CODATA Fundamental Physical Constants site.

A second link, Descriptions, is provided under the NIST Online Databases heading. It also leads to an alphabetical list of all online databases included in the Gateway, but in this case a brief description of each database is also provided. This list is intended to serve the customer who wants to browse before selecting a database. Both these lists are dynamically generated from the content management database.

Similarly, for users who know they want to order a PC database from NIST SRDG, a PC Products link is provided under the Databases for Purchase heading. This link leads to the database price list for all NIST Standard Reference Databases and Special Databases available for purchase.

Users who want to search the Gateway for all databases containing a particular type of data can click on one of the search option links (Keyword, Property, or Substance Name). The options provided with these searches are as follows:

1. Keyword
    a. Enter keyword (asterisks can be used to specify a partial match)
    b. Select keyword from list of all keywords grouped by first letter
2. Property
    a. Property name
        i. Enter property (asterisks can be used to specify a partial match)
        ii. Select property from list of all properties grouped by first letter
    b. Property category
        i. Select major category from list
        ii. Optionally select minor category from list
        iii. Optionally select specific property from list
3. Substance name
    a. Enter substance name
    b. Advanced search (enter up to three partial name strings and specify for each whether the substance name should (1) contain the string, (2) start with the string, (3) end with the string, or (4) not contain the string)
    c. Element (select element from periodic table)

Several of the above search options, namely the keywords and properties grouped by first letter and the property categories, allow the user to select from a list. Each of these lists is dynamically generated from the content management database.

The searches are performed by comparing the search criteria specified by the user against the indexing terms stored in the content management database for the type of search specified (keyword, property, or substance name). Because users may differ in how they interpret the term *keyword*, keyword searches are performed successively against each type of indexing terms until at least one match is found. In other words, if a user enters a keyword and no match is found among the Gateway keywords, the system will search the Gateway properties. Likewise, if no match is found among the Gateway properties, the system will search the Gateway substance names.

All successful searches of the Gateway return a list of links to all NIST databases included in the Gateway that match the search criteria. These links fall into one of two categories:

1. Links to online databases with data matching the search specifications. In this case, the user can select one of these online databases by clicking on its link and then search for the data of interest by following the search menus provided by the selected database. (There is one exception in that the results of some Element searches lead directly to the data of interest. See below for details.) All the online databases are free.
2. Links to web pages of ordering information for PC databases. In this case, the user can select one of these PC databases by clicking on the link to the web page that describes it and provides the option of ordering it. Only a few PC databases are free; most are available for purchase. There are two buttons provided, ONLINE PURCHASE and FAX OR MAIL ORDER. If the user chooses ONLINE PURCHASE, a form is displayed that allows the user to enter a credit card number and order the database electronically. Many of the PC databases can be downloaded and installed on the user's local PC immediately after purchase. After the user either downloads the database or receives it through the mail and installs it locally, the user can run it to search for the data of interest.

The Element search, which is available as an option under the Substance Search, allows the user to select an element from a periodic table as shown in Figure 2.

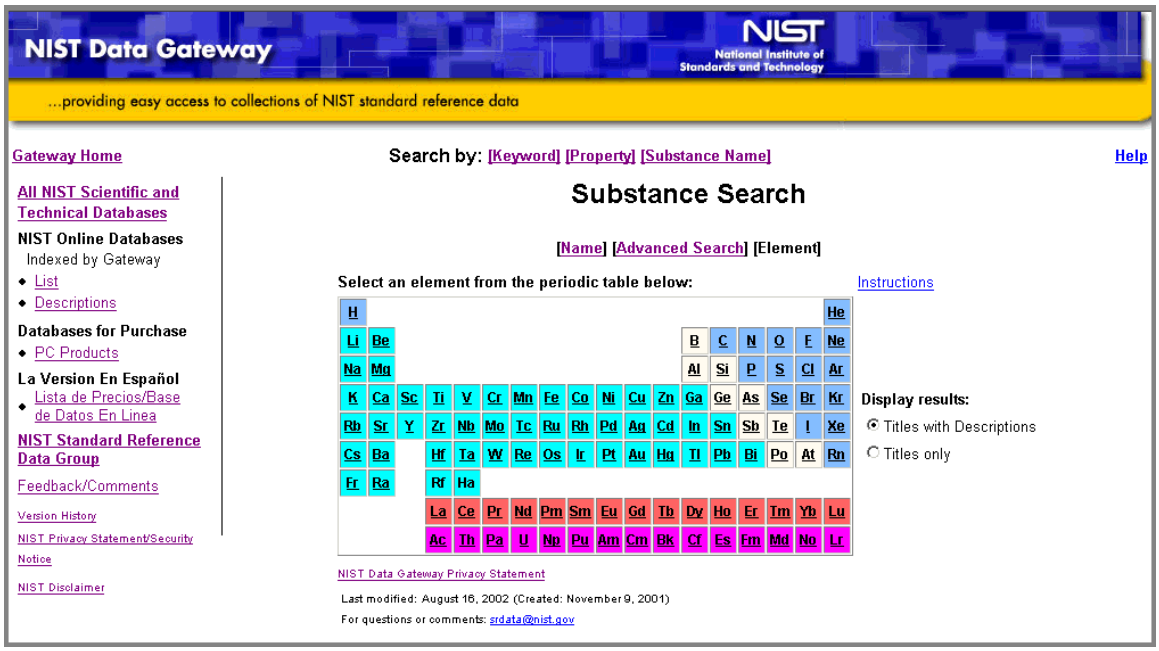

Figure 2. NIST Data Gateway Substance Search for Element

When a user clicks on one of the element links in the periodic table, a list of all databases indexed by that element is presented. The user can choose one of these databases by

clicking on its link. For nine of the online databases in the Gateway system, clicking on the name of the database leads directly to the data page in the database that pertains to the requested element.

Two different options are provided for Gateway users who want to send NIST comments or questions. First, there is a "mailto" link, srdata@nist.gov, for sending email to NIST SRDG. Second, there is a Feedback/Comments link, which brings up a form for users to fill out and submit. This second option allows users to submit comments anonymously, if desired. It also has the advantage that it does not require users to have access to an email program.

Before the Gateway was released, it was carefully reviewed to verify compliance with all U.S. Department of Commerce and NIST policies for web sites. For instance, all necessary modifications were made to ensure that the site met the requirement of being accessible to users with disabilities. In addition, extra code was added so that users with JavaScript [1] disabled in their browsers would have access to functionality equivalent to that provided with JavaScript.

Version 1.0 of the Gateway was originally implemented using frames. The identifying banner, the left column containing the list of common links, and the main portion of the page were each displayed in a different frame. Using the frames technology had the advantage of preserving the search terms entered by the user until the entry box was reset. For example, a user could specify three different criteria for an advanced substance name search and then return to the search screen to change one of the criteria while retaining the other two. However, there are issues associated with creating bookmarks for sites that use frames. For instance, if a user chooses to add a bookmark to his browser while in a frames site, the bookmark may be only for an individual frame and not for the complete web page as displayed. Also, some users have difficulty printing entire screens from a frames site. Therefore, it was decided to eliminate the frames. The Gateway was reprogrammed to function without frames, while still keeping the same design for each page. This "frameless" version was released in April 2002.

2.3 System Architecture

The Gateway was developed using the following three-tier web application architecture:

1. Tier 1 – Presentation layer – HTML
2. Tier 2 – Application layer – Java servlets [1]
3. Tier 3 – Data layer – Content management database and stored procedures in SQL Server [1]

The Gateway web pages are coded using Hypertext Markup Language (HTML). The application is programmed in Java Servlets. JRun 3.1 [1] is used as the Java application server [1]. The content management database is stored in SQL Server 7.0, a relational database management system (RDBMS), along with the stored procedures that were written to search and retrieve data from the database. JDBC-ODBC Bridge is used as the

connection between the Java Servlets in Tier 2 and the SQL Server database in Tier 3. (JDBC and ODBC are acronyms for Java Database Connectivity and Open Database Connectivity, respectively.) Both the JRun application server and the SQL Server RDBMS are located on a server running Windows NT 4.0 [1] as the operating system and Internet Information Server (IIS) 4.0 [1] as the web server as shown in Figure 3.

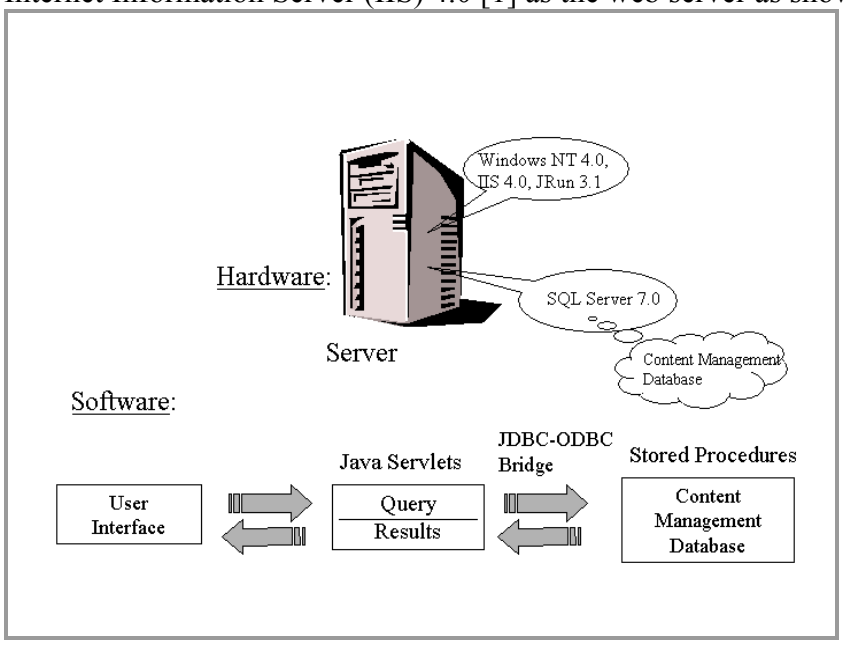

Figure 3. NIST Data Gateway System Architecture

3. Administering the NIST Data Gateway

3.1 Collecting Web Statistics

A Gateway administrative web site was built and implemented concurrently with Version 1.0 of the Gateway. This site is maintained to provide statistics on Gateway usage, including how often different search terms are submitted. These statistics are collected so that they can be used to determine what types of data are desired and which are of most and least interest in order to help NIST serve its customers better. The collection policy is explained to users in the NIST Data Gateway Privacy Statement, which can be accessed from a link near each Gateway submit search button.

A database is maintained of all keywords, properties, and substance names that are submitted as search terms by users and the frequency of each. Records are kept of both successful and unsuccessful (failed) search terms. No personal identifiers are saved. In addition, commonly used web statistics, such as the dates and times users access the Gateway site and the pages they visit, are maintained. A link (NIST Privacy Statement/Security Notice) is provided with the links in the left column of each Gateway page so users can access the complete NIST policy.

The statistics displayed by the administrative web site on successful and failed searches are updated in real time. The remaining statistics are updated daily.

3.2 Evaluating Web Statistics

Now that the Gateway has been operational for more than seven months, some interesting patterns of use are becoming apparent. For example, the keyword search is the most popular. Looking at the statistics on the searches performed in the first seven full months of operation, 52 % of the searches were keyword searches, while 17 % were property searches and 30 % were substance name searches (rounding off to the nearest percentages). Also, users most often chose to type in their keyword search term rather than selecting it from a list. See Table 1 for a summary of the types of searches performed.

	Keyword	Property	Substance name
Type search term or select from list*	6225	2039	3568
Select from property category list		44	
Perform advanced substance search			139**
Select element from periodic table			62
Total	6225	2083	3630

Table 1. Summary of Types of Searches Performed (12/1/2001 to 6/30/2002)

Useful information can be derived from the statistics for the failed search terms, that is, the search queries for which no Gateway databases were found. In some cases, users entered a different form of a word or phrase from what was indexed in the Gateway. For example, *thermal conductivity* is an index term in the Gateway, but not *conductivity*, which showed up as a failed search term. Similarly, *bond dissociation energy* is an index term in the Gateway, but not *bond energy*, which also showed up as a failed search term. A short-term solution can be to add the more commonly failed search terms that should have been successful, but failed simply because they were not included in the content management database for Version 1.0. In the longer term, enhanced searching options need to be investigated. For instance, providing an option to find "similar" terms would be useful, as would incorporating spell checking software and connecting to chemical dictionaries and thesauri.

* The option to select from list is implemented differently depending on the search type. For keyword or property searches, the list is composed of all keywords or properties, respectively, beginning with a user-specified letter. For substance name searches, the list is the results list from an advanced substance search.
** This figure does not contribute to the total number of substance searches. After a user performs an advanced search and selects a substance name from the results list, the search for that substance name is logged with the searches for substance names that are typed into the entry box.

In some cases, search terms failed because no databases containing that type of data were included in the Gateway. One of the most frequently entered failed keywords was *time*. For that reason, it is planned to include the NIST time site, The Official U.S. Time, indexed by *time*, in the next release of the Gateway. It would also be useful to provide users whose searches failed an option to search all NIST webspace. For example, another failed keyword was *fips*. Although there are no Federal Information Processing Standards Publications (FIPS PUBS) in the Gateway, there are numerous references to FIPS at NIST. A search of NIST webspace for *fips* on July 10, 2002, resulted in over 500 hits.

4. Next Steps

An effort is underway to add additional index terms to the content management database based on the statistics collected on failed search terms. The metadata entries for each database included in Version 1.0 are being checked and modifications made as needed to reflect updated versions. Also, metadata are being collected for approximately ten additional databases so they can be included in the next version of the Gateway. In future versions, additional databases will be added on an incremental basis until as many NIST scientific and technical databases as possible are included in the Gateway.

Other enhancements are being considered for future versions. The content management database could be expanded by adding additional indexing terms for the databases included in Version 1.0. For example, substance names and synonyms could be added for the NIST/EPA/NIH Mass Spectral Library, NIST '98 Database ASCII version and the NIST/EPA/NIH Mass Spectral Library, NIST '98 With Windows 2.0 Software. (These databases were indexed only by keyword and property for Version 1.0 of the Gateway.) Links to additional NIST data resources could be added to the Gateway interface as more resources become available. For example, a link to a comprehensive list of NIST databases could be provided in the left column of the Gateway pages.

A challenge for the longer term will be to expand the concept of the Gateway so it can serve as a tool for integrating various NIST data resources. For example, an effort will be made to integrate some of the NIST online systems more fully into the Gateway by providing users with functionality that extends beyond just providing links and then passing control to the requested systems.

5. Conclusion

The NIST Data Gateway, a web portal built to provide easy access to NIST scientific and technical data resources, was launched in November 2001. Seventy-four NIST online and PC databases are indexed in Version 1.0 of the Gateway, which allows users to specify keyword, property, or substance name and find the NIST databases that contain standard reference data meeting the specified criteria. The experience gained from the development of the Gateway, as well as the feedback and comments received from reviewers and users, should help NIST as it continues to strive to improve the accessibility of NIST data resources to customers. Future releases of the Gateway are

planned to provide access to additional NIST data resources and to enhance the functionality provided. Finally, the lessons learned throughout the different phases of the Gateway project could be applied to other data portal projects.

6. Acknowledgements

The authors wish to thank John Rumble for the vision he provided for a NIST data portal and for his support of the project, and Joan Fuller for her valuable contributions to the development of the original concept for the Gateway and the prototype. In addition, the authors wish to thank the other Gateway Team members, William Dinis Camara, Geraldine Dalton, Linda Diane Decker, Sherena Gray Johnson, Joan Sauerwein, and Shari Young, for their many valuable contributions. Finally, the authors wish to thank the many people who provided advice and assistance during this project, including Edwin Begley, Donald Burgess, Joseph Conny, Robert Dragoset, Phoebe Fagan, Robert Goldberg, Cheryl Williams Levey, Peter Linstrom, Bijan Mashayekhi, Ronald Munro, Lane Sander, and Han Thai.

7. Notes

[1] Certain trade names and company products are mentioned in the text to specify adequately the computer products needed to develop this data system. In no case does such identification imply endorsement by the National Institute of Standards and Technology of these computer products, nor does it imply that the products are necessarily the best available for the purpose.