Dealing With Data
Seminar Summary

A Summary of Seminar #1 of the
Using Information in Government Program Seminar Series

February 4, 1999

Center for Technology in Government
University at Albany / SUNY

The Center for Technology in Government, through the Using Information in Government (UIG) program, has worked during the past year with NYS agency project teams and partners from the public, private, and academic sectors to identify benefits of and strategies for integrating and using data for program planning, evaluation, and decision making. The policy, management, and technology issues identified through our work with the agency teams will be shared in a series of seminars focused on increasing the value of existing information to government programs. This report summarizes the presentations given at the first of the Using Information in Government seminar series, "Dealing with Data" conducted on February 4, 1999.

The first seminar covered a wide range of data issues such as quality, standards, and long-term maintenance and preservation. The seminar was divided into three sessions. In the introductory session, Dr. Giri Kumar Tayi, professor of Management Science and Information Systems at the University at Albany, gave a presentation on data quality management. In the second session, three speakers- Wendy Scheening, Manager of Data Processing Technical Services at the NYS Department of Agriculture and Markets; Pam Neely, Graduate Assistant at the Center for Technology in Government and doctoral student in the Information Science program at the Nelson A. Rockefeller College of Public Affairs and Policy of SUNY Albany; and Michael Medvesky, Director of the Public Health Information Group at the NYS Department of Health- presented data tools and techniques. Finally, Alex Roberts, Assistant Director of Data Processing at the NYS Division of Criminal Justice Services, and Alan Kowlowitz, Manager of Electronic Records Management Services at the NYS Archives and Records Administration, shared their experience in dealing with data issues. A panel discussion concluded the program.

Welcome

Theresa Pardo, Project Director, Center for Technology in Government

Theresa welcomed everyone to the first of the seminar series based on the results of the Using Information in Government Program. She noted that the purpose of the Using Information in Government program is the management of information that has value to government agencies and which is often expensive and difficult to acquire. The objective of the seminar series is to share the lessons learned during Using Information in Government with program and IT managers in various levels of government. She gave a brief overview of the seminars. The first seminar deals with identifying and overcoming various data issues. The next seminar, "Information Use Tools and Skill Sets" which will be held on May 4, 1999, will address public managers' needs as information users. The third seminar, "What Rules Govern the Use of Information?" on October 5, 1999 will deal with policy issues in the field of information sharing. The final seminar, in January 2000, will be a capstone event that will address the full set of lessons learned throughout the two years of the Using Information in Government program.

Theresa provided the background of the Using Information In Government Program. She discussed the issues that public managers face in using government information to do their jobs. The information use issues identified were: (1) a lack of incentive to share information; (2) a lack of understanding of the value of integrating data and using it to support decision making and planning; (3) a lack of understanding of the technical, human, and organizational requirements; and (4) a lack of understanding of the real potential of the technology.

Theresa reviewed the objectives of the Using Information in Government program:

- To recommend policies or policy templates to guide public officials in their use of government information;
- To develop and assess data standards, inventories, and quality assurance tools;
- To develop and assess cost-benefit models and other measures of information value;
- To assess the cost-effectiveness of various technical tools and techniques;
- To develop collaborative and collective resources for data users.

The proposed practical products that will be generated from the program are: (1) practical guidelines for developing an IT business case, (2) the seminar series, (3) a cost performance model for projects that focus on information sharing, and (4) case studies based on the agency projects conducted over the full two-year period. In addition, the Center for Technology in Government will provide feedback to the Office for Technology (OFT) regarding the NYS Information and Technology Policies that are applied in these projects. Theresa also presented a synopsis of the three agency projects completed during the first year of the UIG program: the Office of State Comptroller, Division of Municipal affairs; the Central New York Psychiatric Center; and the Office of Temporary and Disability Assistance, Division of Audit and Quality Control, Bureau of Shelter Services.

Finally, Theresa provided a few insights regarding dealing with data in the public sector. Data quality issues are a major and ongoing concern, which occupy 80% of efforts to use data. Data standards are also key to sharing data across agencies and for integrating data from multiple sources. There are a variety of tools to support efforts to address issues with existing data sources. She mentioned that many efforts to bring together disparate data sources to form new information resources have been successful, but that the road is fraught with risk. Finally, she said that preservation issues must be considered when systems are planned.

Data Quality Issues

Giri Kumar Tayi, Associate Professor, School of Business, Management Science and Information Systems, University at Albany

Giri Kumar Tayi gave a presentation on data quality management tailored to users and consumers of data. The first part of the presentation focused on key ideas and broad concepts around data quality. The second part dealt with data quality in the context of the public sector, and the third part proposed one approach to deliver data quality.

Giri presented an analysis of the similarities and differences of information age and industrial age resources. Information age resources do not consist of labor, capital, raw material, and energy, but of data, information, and knowledge. Traditional and information age resources differ in their very nature. It is important to recognize these differences in order to manage them effectively. Giri described the special characteristics of data. Data is intangible- whereas you can physically touch raw material. Data is not consumable- you can use it over and over again as opposed to any raw material which has a one time use. Data is shareable—you can easily share data across agencies. Data is easy to copy at a low cost—the main cost being to produce it not to copy it. Data can be more easily and automatically transported than raw material. Data is not fungible—you cannot substitute one part for another. Data is very fragile—it can easily be erased as it is magnetic. Data is versatile—it can be used by different parties for different uses. Data does not have market valuation—it is very difficult to come up with an economic value for data. Data is not depreciable—it does not get used up even if you use it a million times. He said that data quality addresses " fitness for use" of the raw material of the information age. You have to ask yourself how fit is the data for use, and for what purpose?

Giri mentioned that you have to be very specific about data quality and think about it in terms of attributes. He presented four broad categories of attributes that are important to manage in order to have good quality of data.

Table 1: Categories and Dimensions of Data Quality

CATEGORIES	ATTRIBUTES
Intrinsic	Accuracy Objectivity Believability Reputation
Contextual	Completeness Timeliness Relevancy Value Added
Representational	Interpretability Ease of Understanding Concise & Consistent representation
Accessibility	Accessibility Access security

Source: Beyond Accuracy: What Data Quality means to Data Consumers. Wang, et al (1994).

The problem with these attributes is that some are visible— such as accuracy, timeliness, usability—and others are not. Users tend to be realistic about visible factors while being

unduly optimistic about the invisible ones. There is major trade-off that takes place between these attributes. You often need to give something up and it is important to make the right trade-off. For example, all things being equal, the more timely the data the better it is. However, very seldom are all things equal. The appropriate trade-off depends on the value of the data at different points of the processing spectrum. You start with a rough piece of data and then value is added over time so the value dimensions of the data keep changing. Data quality management means different things depending on your perspective. The definition of data quality has two perspectives. From the analyst's perspective, data quality management requires a sound understanding of the nature of data, identifying the factors that determine its quality, and articulating the underlying trade-offs. From an organizational perspective, data quality management involves specification of policies, identification of techniques, and use of procedures to ensure that the organizational data resource possesses a level of quality commensurate with the various uses of data. In transforming data into information, individuals play one or more roles. The three main roles are: (1) Data producers-people or groups who generate data; (2) Data custodians- people who provide and manage computing resources for storing and processing data (IT people); and (3) Data consumers- people or groups who use data. Each group has specific processes and tasks that you need to think about.

Giri provided a set of questions that need to be answered in order to assess the quality of data in an organizational context:

1. What is the data element of interest? Is it specific numbers, findings, conclusions?
2. How was the data or information created? When? Understanding the sources of bias and knowing the methods involved in creation of the data are important;
3. Who is associated with the data creation? Different audiences perceive the credibility of the data differently;
4. From what viewpoint was the data created and why? Every data element has a perspective that results from the context of its generation;
5. What relationship does this data element have to other data elements? Discerning the relationship is an important part of assessing data quality;
6. What approval, review and filtering process has the data endured? This is very important in the public sector as it gives believability to the data.

He recommended that approaches to enhance data quality focus on improving the data itself, improving the mechanisms that collect and deliver data, and improving the ability of an individual to assess the data quality for a specific purpose. These goals are not mutually exclusive and should all be pursued.

Giri presented an intrinsic data quality problem pattern developed by Strong, Lee and Wang (1997). It starts with having multiple sources of the same data. As a consequence, mismatches occur. Because of these mismatches, believability becomes questionable. The application becomes sloppy or the organization uses it only partially. The poor intrinsic data quality becomes common knowledge. Therefore, the data may end up not being used because of little added value and poor reputation. Mismatches among sources of the same data are a common cause of intrinsic data quality concerns.

Giri addressed the issue of how to deliver high quality data or information. The problem is that in most organizations, data or information is managed as the by-product of a system or an event. However, the consumer or user of the data views it as a product, not a by-product. Organizations often focus exclusively on the hardware or software components of the system rather than the data. Moreover, these components are managed in isolation and as a result the means of producing information becomes an end in itself. Giri presented a four-step approach to delivering high quality data developed by Wang et al. (1998). The first step is to understand the consumer needs. The goal is to ensure that the data is fit for consumer use. It is a total product that exhibits all the attributes that meet or exceed the consumer' s expectations. You need to look at all the dimensions: timeliness, accuracy, etc. The second step is to manage the process. The process must be well defined and must contain adequate controls, inspection and production, and delivery time management. The third step is to manage the life cycle of the data. Just as in the case of a physical product, data products should be managed over their entire life cycle, keeping in mind the nature of the data, the tasks it supports, and the changing environment in which it is used. The last step is to delegate the responsibility of data quality to a single individual. He or she could coordinate and manage the three key stakeholder groups: suppliers of raw data, producers of the deliverable data product, and consumers of the data product. Giri characterized this as an integrated, cross-functional approach.

Finally, Giri provided a set of some general data quality rules developed by Orr (1996):

- Data that is not used cannot be correct for very long.
- Data quality in an information system is a function of its use, not its collection.
- Data quality will, ultimately, be no better than its most stringent use.
- Data quality problems tend to become worse with the age of the system.
- The less likely some data attribute (element) is to change, the more traumatic it will be when it finally does change.
- Laws of data quality apply equally to data and meta data.
- Variations among the data sources' attitudes, policies, and practices contribute to uneven data quality.

Statewide Data Standards: What Can They Do For You?

Wendy Scheening, Manager of Data Processing Technical Services,
NYS Department of Agriculture and Markets

Wendy Scheening gave a presentation on New York State' s efforts to establish common data standards. First she stated that information is an asset which allows us to make more informed decisions to provide better governance. However, it is important to know that not all data is an asset. Indeed, not all data leads to information, not all data is useful, and information overload can be harmful. It is important to discern what is relevant among the huge amount of data and information available. She insisted that only "good" data is an asset. She defined "good" data as having the following characteristics: (1) specific business relevance; (2) common understanding between business partners and agreement

on how the data is collected; (3) concise semantic definition; (4) completeness (e.g. an empty field might be confusing); (5) appropriate values (e.g. the date 9999 can be confusing as 9 is often used for missing values); and (6) leading to information and knowledge.

Wendy then presented reasons why standards are useful. First, they help create "good data" as they increase consistency and improve validity. For example, they reduce the questions one may have on how the information was collected. Second, they facilitate the use of data, by allowing data sharing, for example. Third, they foster a common understanding among business partners and promote semantic clarity.

Some additional reasons why standards are beneficial are:

- Standards facilitate communications between government agencies by having common data definitions for electronic data interchange and/or shared databases.
- They improve management decisions by simplifying the integration required to bring a variety of data sources together.
- They enable the integration of systems by enabling agencies to co-develop and reuse databases and programming modules that support common cross-agency functions.

Wendy presented a few practical examples concerning semantic clarity, common coded values (ex: county code), data matching, year 2000, multi-partner data exchange, and joint application development.

Finally, she gave a few characteristics of data standards:

- Data standards consist of a dictionary framework and preferred standards,
- They accommodate data exchange and storage,
- They allow a variety of data designs and structures,
- They are non-proprietary, and they are an evolving project.

She concluded by saying that statewide data standards are good for New York. She also provided a set of references for more in-depth information about data standards:

Office for Technology Web Site:
http://www.irm.state.ny.us/

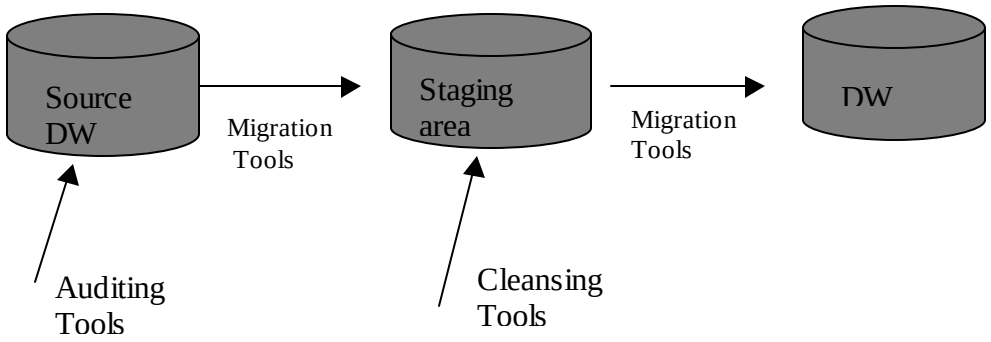
ICEDP Web Site
http://www.icedp.org/

NYS Forum for Information Resource Management
http://www.nysfirm.org/

Data Quality Tools

Pam Neely, Graduate Assistant, Center for Technology in Government

Pam Neely gave a presentation on how data problems can be mapped to features of data quality tools. She began by explaining the basic process of building a data warehouse. Data usually comes from one or several source databases, migrates to an intermediate storage area where the data is transformed, then it migrates to the data warehouse. Along the way, there are tools—auditing tools, cleansing tools and migration tools—to help ensure that the data is clean and accurate prior to being used to populate the data warehouse:



Auditing tools are used at the source, cleansing tools are used in the intermediate stage, and migration tools are used to move the data from the source to the staging area and then to the data warehouse. The data auditing tools ensure accuracy and correctness at the source of the data; they also compare the data to a set of business rules to perform validation checks. Data cleansing tools are used in the intermediate stage. They cleanse data by breaking the records into atomic units, standardizing the elements, matching records against each other to check for duplication, and consolidating data. Finally, data migration tools extract data from a source database, move it to the intermediate staging area, map data to the data warehouse schema, and move it into the warehouse. Pam presented a matrix mapping data quality problems to a selected set of data quality tools.

Table 2: Data Quality Matrix

Problems	**Features**	**Tools**
Auditing Tools		
Is your data complete and valid?	Data examination- determines quality of data, patterns within it, and number of different fields used	WizSoft- WizRule Vality- Integrity
How well does the data reflect the business rules? Do you have missing values, illegal values, inconsistent values, invalid relationships?	Compare to business rules and Assess data for consistency and completeness against rules	Prism Solutions, Inc.- Prism Quality Manager WizSoft - WizRule Vality- Integrity
Are you using sources that do not comply to your business rules?	Data reengineering- examining the data to determine what the business rules are	WizSoft – WizRule Vality- Integrity
Cleansing Tools		
Does your data need to be broken up between source and data warehouse?	Data parsing (elementizing)- context and destination of each component of each field	Trillium Software- Parser i.d. Centric- DataRight
Does your data have abbreviations that should be changed to insure consistency?	Data standardizing- converting data elements to forms that are standard throughout the DW	Trillium Software- Parser i.d. Centric- DataRight

Is your data correct?	Data correction and verification- matches data against known lists (addresses, product lists, customer lists)	Trillium Software- Parser Trillium Software- GeoCoder i.d. Centric- ACE, Clear I.D. Library Group 1- NADIS
Is there redundancy in your data?	Record matching- determines whether two records represent data on the same object	Trillium Software- Matcher Innovative Systems- Match i.d. Centric- Match/Consolidation Group 1- Merge/Purge Plus
Are there mutiple versions of company names in your database?	Record matching- based on user specified fields such as tax ID	Innovative Systems- Corp-Match
Is your data consistent prior to entering data warehouse?	Transform data- "1" for male, "2" for female becomes "M" & "F" - ensures consistent mapping between source systems and data warehouse	Vality- Integrity i.d. Centric- Match/Consolidation
Do you have information in free form fields that differs between databases?	Data reengineering- examining the data to determine what the business rules are	Vality- Integrity
Do you multiple individuals in the same household that need to be grouped together?	Householding- combining individual records that have same address	i.d. Centric- Match/Consolidation Trillium Software- Matcher
Does your data contain atypical words- such as industry specific words, ethnic or hyphenated names?	Data parsing combined with data verification- comparison to industry specific lists	i.d. Centric- ACE, Clear I.D.
Migration and Other Tools		
Do you have multiple formats to be accessed- relational dbs, flat files, etc.?	Access the data then map it to the dw schema	Enterprise/Integrator by Carleton.
Do you have free form text that needs to be indexed, classified, other?	Text mining- extracts meaning and relevance from large amounts of information	Semio- SemioMap
Have the rules established during the data cleansing steps been reflected in the metadata?	Documenting- documenting the results of the data cleansing steps in the metadata	
Is data Y2K compliant?		
Is the quality of the data poor and people don' t care because they have adjusted to it?		

Finally, Pam mentioned a few important questions that data quality tools cannot address:
(1) Have the users of the source database adjusted to poor quality of the database and developed "work-arounds"?
(2) Is there conflicting information that can't be compared to a known resource?
(3) Do you have a lot of soft data that needs to be placed in your data warehouse?

Using Data Tools in the Health Information Network

Michael Medvesky, Director, Public Health Information Group, NYS Department of Health

Michael Medvesky gave a presentation on the NYS Health Information Network (HIN). The Health Information Network is a secure intranet system for use by public health officials, county health directors, and the NYS Department of Health (DOH). It is an excellent communication tool, allowing the safe exchange, sharing, and submission of

data. In addition, the HIN assures local health departments timely and secure access to queriable data sets such as the Statewide Planning and Research Cooperative System (SPARCS) data, Tuberculosis data and Communicable Disease registry data, as well as other relevant health information, documents, reports, press releases, and products that can be used for community health assessment and planning. The unique feature of the HIN is a Web server technology that provides a standardized and secure environment for entering data and accessing information resources for both local and state. The environment also eliminates the expensive processes of distributing and maintaining software programs, providing for improved efficiency and productivity for both state and county staff.

The NYSDOH HIN was developed with funding assistance from the Center for Disease Control and an INPHO grant. The HIN project received the NYS Forum for Information Resource Management' s Best Practice Award in 1996. The US Center for Disease Control (CDC) has also cited the HIN project as a model for other state and federal Health Information Networks.

Mike presented some of the data issues associated with the HIN. Regarding the use and misuse of information, they had to: (1) address timeliness of information; (2) aggregate population estimates at county and subcounty level; (3) determine crude versus adjusted rates (what standard population to use?); and (4) deal with small area analysis. Some data access issues that they had to address were that the HIN is available at the local level but people outside the local health department cannot access it. Providing access to those who need it, helping them get access, and keeping out others who should not have access are important considerations. Finally, staff support is also an issue as HIN users require technical assistance, as well as update and maintenance support.

Experience with a Large Database Redesign and Conversion

Alex Roberts, Assistant Director of Data Processing,
NYS Division of Criminal Justice Services

Alex Roberts talked about the Division of Criminal Justice Services (DCJS) experience in redesigning and converting their Computerized Criminal History (CCH) system. This large database contains 9.1 million histories that are made up of 13 million criminal cycles. It has a growth of an average of 3000 histories/day. Therefore, the project is mainly a computer systems migration project, involving the redesign of a very large mainframe-based, on line system to one that uses object-oriented design, 3-tier client/server architecture and relational database design.

The major principles that they applied to the data design were the following:

1. Storing all data in standard format unless there is a technical impediment to doing so. In that case, the data would be stored in a way that can ensure data standards transmission compliance.

2. Translating DCJS codes into standard codes and eliminate the DCJS codes. When the data standard elements are inadequate, they would be stored in a format that is conducive to good data modeling principles. Where there are conflicting standards (criminal justice versus statewide), the standard that is most beneficial would be used. An ability to transmit in either standard is required.
3. Documenting all changes to the way current data are stored. If the decision were made not to store in data standards, the reasons would also be documented.

The problems that they encountered dealt with the conversion of data to comply with established data standards; the conversion of invalid data, and the conversion of incorrect data (data that passes constraint edits, but is not correct). The successes were the following: (1) development and adherence to a data standards policy; (2) only one element (DOB) with invalid dates was converted to new database using 2 data elements; (3) resources were allocated to clean up the most serious problems in the conversion of incorrect data, mainly by researching source documents.

The timeline of the project:
- 2 years of database design, and redesign,
- 1 year of analysis and coding of the conversion program,
- multiple cycles of test conversions,
- projected 4 months conversion run,
- the 2 databases must continue to run in parallel for several years while the migration project is completed.

Alex ended his presentation by giving a few recommendations:
- Develop a thorough understanding of the data. Review the current processing. Talk to people who use the data;
- Don't be seduced by development speed. Take time and care in the analysis, design, testing, and tuning;
- Never misuse data; each data element should have only one definition.

Preserving Information in Government

Alan Kowlowitz, Manager, Electronic Records Management Services
NYS Archives and Records Administration

Alan Kowlowitz gave a presentation on data preservation in government agencies. He started by listing a series of problems in data preservation:
- Preserving access and usability over time requires ongoing maintenance of the data, which can be costly.
- The costs are acceptable for information with immediate business value.
- The issue of preservation is usually not addressed when systems are planned.
- The information without immediate business value is at risk.

He then gave a few examples of information resources lost due to lack of attention to long term preservation issues. He presented two examples at the federal level: the 1960 Census and the National Aeronautical and Space Agency (NASA) information that had been stored on equipment no longer available. At the state level, in New York, he gave the examples of the Land Use National Resource System (LUNRS), the Children Youth Management Information System (CYIMIS), and the Committee on Sentencing Guidelines. Information generated by all three programs is no longer usable.

He presented a series of reasons for data preservation:
(1) a clear legal requirement, it is more easy if the agency has a mandate;
(2) a long-term programmatic need, i.e. if the agency see a long-term use;
(3) existence of secondary users, it increases the value of information and the risk of managed information;
(4) high risk and visibility, agencies want to be able to answer questions;
(5) demonstrable benefits; and
(6) organizational culture.

Alan gave a number of examples of organizations that have successfully addressed their data preservation issues: DOH Vital Records/Disease Registries, DCJS Trends Data, DED BEDS, and DOCS Under-Custody. The future direction for SARA is to address data preservation issues, specifically how to reduce costs through the use of more focused system planning, technology, organizational models, and education.

Panel Discussion

Panel members: Alan Kowlowitz, Michael Medvesky, Alex Roberts, Wendy Scheening, Giri Kumar Tayi; Panel discussion moderator: Sharon Dawes, Director, Center for Technology in Government

Sharon Dawes asked the panel: "When you acquire information from another organization for use by your agency, how do you go about determining its quality and suitability for your use?"

Michael Medvesky said that he would follow the approach that Giri Tayi presented in his session "Data Quality Issues." He would first look at the context, determine why his agency wants this information and what they will use it for. Then he would look at the background of the information, try to find out why the data was collected in the first place. Finally, he would look at data quality issues, and investigate how the collecting agency assessed data quality for its own purposes.

Giri Tayi gave an analogy of a manufacturer going to his supplier to check if the products are fit to his needs. He said that the same relationship needs to exist between the supplier and user of data. Thinking of this analogy can provide a good model to determine if the data is suitable for use. Alan Kowlowitz mentioned that with the use of the Internet, almost all state agencies are in the public access. It would be useful to explain the

context of data creation to the population of users. Wendy Scheening added that if you do not know where the data comes from or what its limitations are, it might be misused.

Sharon Dawes then asked the panel "If you had one piece of advice to give to people grappling with data issues, what would it be?"

- Wendy Scheening replied "try, try, try!" She said the move towards standardization is a long, slow, frustrating process and it is often hard to see the payoff at the beginning, but she believes it does pay off in the long run.
- Giri Tayi said that trying to ensure a high level of data quality in organizations is a journey, an ongoing process that requires continuous attention.
- Alan mentioned that issues have to be addressed at all key milestones in a project or system life cycle.
- Alex stressed the critical importance of the staff understanding of the business use of the information, not just its technical characteristics.
- Mike mentioned having a process for building systems that meet the needs of all users.

Closing

Sharon Dawes

Sharon Dawes concluded the seminar by summarizing what had been learned during the day. She reiterated that information is cheap, but that relevant information is very expensive and hard to get. The presentation by Giri Tayi explained why this is true and other speakers provided additional information and an experience base of practical examples and lessons. The common thread of the presentations was that dealing with data issues is a journey: you need to be deliberate about the journey; you need to emphasize the business use; look at a variety of uses; and be persistent in pursuing data quality tools.

Center for Technology in Government
University at Albany / SUNY
1535 Western Avenue
Albany, NY 12203
Phone: 518-442-3892
Fax: 518-442-3886
info@ctg.albany.edu
www.ctg.albany.edu