WP. 7
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Geneva, Switzerland, 9-11 November 2005)

Topic (i): Web/on-line remote access

ANalytical Data Research by Email and Web (ANDREW)

Supporting Paper

Submitted by the National Center for Health Statistics, Centers for Disease Control and Prevention, United States of America[1]

[1] Prepared by Vijay Gambhir and Kenneth W. Harris (rdca@cdc.gov).

10/03/2005

ANalytical Data Research by Email and Web (ANDREW)

Vijay Gambhir and Kenneth W. Harris, Research Data Center, National Center for Health Statistics, Centers for Disease Control and Prevention, Hyattsville, MD 20782, USA, rdca@cdc.gov

Abstract: The NCHS Research Data Center (RDC), established in 1998, is a facility at the NCHS headquarters in Hyattsville, Maryland, where researchers are granted access to restricted data files, in a secure environment, that are needed to complete approved projects. Restricted data files may contain information, such as lower levels of geography, but do not contain direct identifiers (e.g., name or social security number). Identifiable data include not only direct identifiers such as name, social security number, etc., but also data that can serve to allow inferential identification of either individual or institutional respondents by a number of means.

Although it was envisioned that most of the research work would be performed by the data analysts onsite, with RDC staff closely monitoring to assure that confidential or identifiable data do not leave RDC premises, remote access capabilities were considered an integral part of the fundamental set up of RDC. Thus, the software engineers at RDC designed and developed an e-mail based remote access system, **AN**alytical **D**ata **R**esearch by **E-**mail (**ANDRE)**.

The main objective of ANDRE is to provide a convenient, reliable, economical, and flexible tool for remote data access for statistical analysis. Although ANDRE has served the data users' community very well while strictly adhering to the confidentiality restrictions of NCHS, it does have certain limitations and constraints inherent in its design. Most of these constraints are non-critical but make the system less flexible and less efficient than onsite analysis. Some of these constraints were known at the design and development stage of ANDRE; others have been compiled by RDC staff from interaction with the users and from several years of regular performance analysis of the system. As a result, the RDC now plans to develop a new remote access system, **AN**alytical **D**ata **R**esearch by **E**-mail

and **W**eb (**ANDREW)**. This new system will address research needs of data analysts at all levels. It will support multiple statistical languages (SAS, Sudaan, and Stata) and will provide a Graphic User Interface (GUI) for language free statistical analysis. In addition, the system will address the problem of confidentiality risks resulting from cumulative data retrieval through multiple requests from the same user.

1 RDC and data access for statistical analysis

Despite the wide dissemination of its data through publications, CD-ROMs, etc., the inability to release files with, for instance, lower levels of geography, severely limits the utility of some data for research, policy, and programmatic purposes and sets a boundary on one of the Center's goals to increase its capacity to provide state and local area estimates. In pursuit of this goal and in response to the research community's interest in restricted data, NCHS established the Research Data Center (RDC), a mechanism whereby researchers can access detailed data files in a secure environment, without jeopardizing the confidentiality of the respondents.

The NCHS Research Data Center, established in 1998, is a facility at the NCHS headquarters in Hyattsville, Maryland, where researchers are granted access to restricted data files needed to complete approved projects. Restricted data files may contain information, such as lower levels of geography, but do not contain direct identifiers (e.g., name or social security number). Identifiable data includes not only direct identifiers such as name, social security number, etc., but also data that can serve to allow inferential identification of either individual or institutional respondents by a number of means.

2 Remote Access/ANalytical Data Research by Email (ANDRE)

Although it was envisioned that most of the research work would be performed by the data analysts onsite, with RDC staff closely monitoring to assure that confidential or identifiable data do not leave RDC premises, remote access capabilities were considered an integral part of the fundamental set up of RDC. The software engineers at RDC designed and developed an email based remote access system that enables researchers to perform **AN**alytical **D**ata **R**esearch by **E**mail. Thus the system was named ANDRE.

The main objective of ANDRE is to provide a convenient, reliable, economical, and flexible tool for remote data access for statistical analysis. Since 1998 ANDRE has served many data analysts flawlessly. A number of researchers have used ANDRE in conjunction with onsite sessions either performing preliminary analysis before onsite sessions or performing post onsite session analysis to wrap up their research. However, most of the analysts have used it for independent statistical research.

3 ANalytical Data Research by Email and Web (ANDREW):

Although ANDRE, the existing remote access system, has served the data users' community very well while strictly adhering to the confidentiality restrictions of NCHS, it does have certain limitations and constraints inherent in its design. This new remote access system will address research needs of data analysts at all levels. It will be built upon the time-tested architecture of its predecessor with a few enhancements to its algorithms. Also, it will incorporate new features to make it robust, very strong on confidentiality issues, more efficient and more flexible.

3.1 Basic Layout:

ANDREW will be a fully automated remote access system that will serve registered users around the clock without human intervention. To subscribe to the system, a data user will be required to submit a research proposal. Once the proposal is approved, the user will be provided with login information and guidance as to how to use the system. A registered user may submit data requests from anywhere and at any time. However, results of the data requests will always be released to a specific email address that has been certified as secure and approved by RDC.

The system will use a multilayered authentication procedure to ward off unauthorized access. Upon receipt of a data request, ANDREW will verify login

credentials and subscription status of the requester. Unauthorized communications will be discarded without any response.

A user's program from a validated data request will be scanned online for its suitability for execution by ANDREW. To ensure smooth operations, it will not allow certain commands and words in a user's programs, especially those that can create permanent datasets or files on ANDREW's disk space or interfere in any way with the underlying operating system.

To deal with the issues of disclosure limitation, ANDREW will use prevention as well as suppression techniques. To prevent disclosure violations, it will not allow certain commands (e.g., print command in SAS) that have little, if any, statistical value. Also, it will modify certain commands in the user's program to prevent output of sensitive information. For example, it will modify the "proc means" command so that it does not produce minimum and maximum values. However, there are certain commands that are important for statistical studies and do generate output that cannot be released. ANDREW will use a variety of enhanced suppression algorithms to prevent disclosure violations. For example, it will white out extreme values resulting from proc univariate and will use state of the art commercial software packages to suppress certain low values in one-way and two-way tables with a special emphasis on prevention of inferential disclosure violations.

3.2 New Features:

3.2.1 Data Security:

The technological platform used for the development of ANDRE is almost obsolete and is prone to attack by hackers. ANDREW will be designed using MS Visual Studio. The GUI front end will be implemented using C# (C Sharp) and the backend will be a product from a leading corporation which has proven technical know-how as well as commitment and resources to keep the platform secure by issuing security patch ups to keep the product safe.

3.2.2 Robust disclosure limitation checks

Although the suppression algorithms employed by ANDRE have performed well, certain limitations and constraints have made the system less effective. For example, ANDRE works directly on the SAS output (.lst files) and has to

negotiate labels, formatting characters, and background information in order to evaluate statistical values and apply home grown suppression algorithms. By contrast, ANDREW will develop a set of mapping algorithms that will extract values from the SAS/Sudaan/Stata outputs and save them in external files in an appropriate format without any background information. The system will invoke a well recognized commercial suppression software package. Another set of algorithms will put the validated values back into the original SAS/Sudaan/Stata output.

3.2.3 Cumulative data retrieval and confidentiality

No solution has yet been found to the classic problem of a user submitting a series of data requests through a remote access system, each time getting different bits and pieces of data and eventually getting sufficient information to identify entities in the data set. However, ANDREW will start addressing this problem by tracking the amount and type of data released about certain risky/critical variables. A committee of confidentiality experts will examine each data set and identify variables that have disclosure violation potential. For each variable a tolerance level will be defined. ANDREW will examine data being released for each risky variable. As soon as the amount of data released for any variable reaches its tolerance level, ANDREW will issue an alert to the system administrator and generate a report displaying all the data releases for that variable.

3.2.4 Accessibility

Since the user interface of ANDREW will have no confidentiality risks, it will be a Web based component. This feature will not only give wider accessibility to the system but will also allow a user to get his/her program parsed interactively for suitability before it is accepted for execution by ANDREW. However, all of the confidential data along with the main resident component of ANDREW will reside on a set of machines physically located in RDC's secure area. The main resident component of the system will receive the parsed and validated users' programs from the web component, execute them in the secure environment and send the results to the appropriate users.

3.2.5 Multi-language support

Unlike SAS, other statistical languages such as Sudaan produce a very rich variety of output patterns. Remote access systems rely heavily on pattern recognition algorithms to identify disclosure violations. With older technology matching is done at the character level, whereas C# has a built in feature called

regular expressions that can detect a variety of patterns at a block level rather quickly. This will allow the system to deal with all kinds of patterns generated by Sudaan and other languages in a timely fashion.

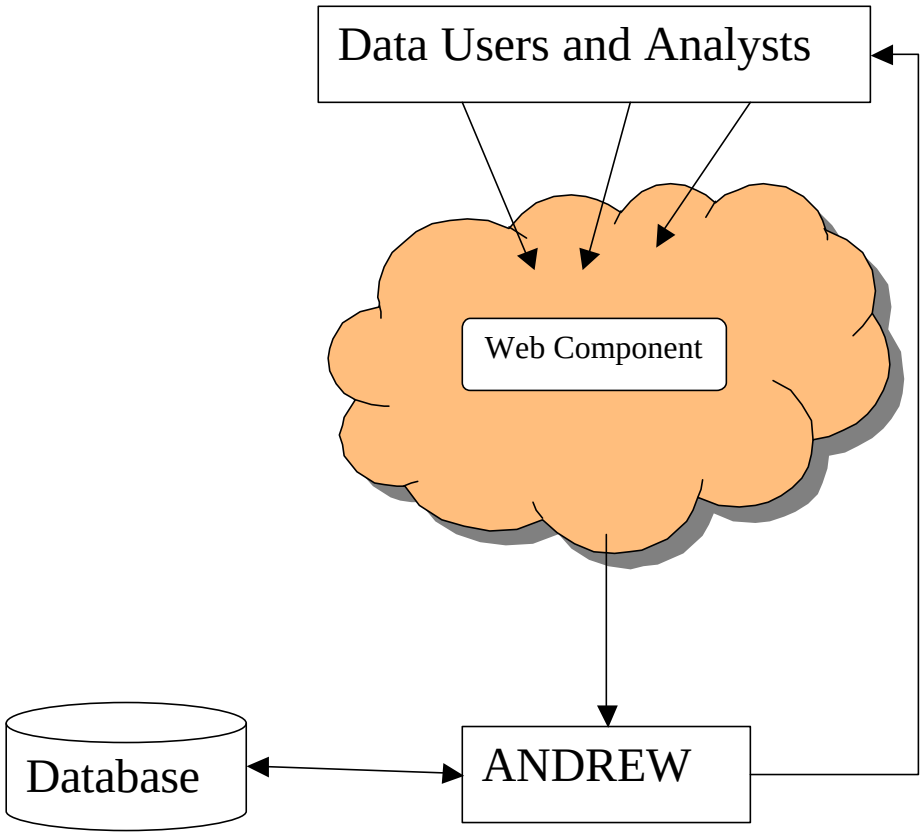
3.2.6 Language free GUI data access

This feature of ANDREW will implement a Graphic User Interface (GUI) that will allow specification of desired variables and the constraints by a few mouse clicks. GUI will be very useful to the users whose SAS skills are rusty/non-existent or who want to get quick results. Use of GUI technology has futuristic implications as it will give a lot more control on what a user can specify compared with the current approach of giving free hand to a user (via his or her SAS code) once the data set is approved. The confidential data will also reside in RDC's secure area and appropriate SAS/Sudaan/Stata codes will be generated and executed in the secure environment.

4 Graphic Overview:

4.1 The overview:

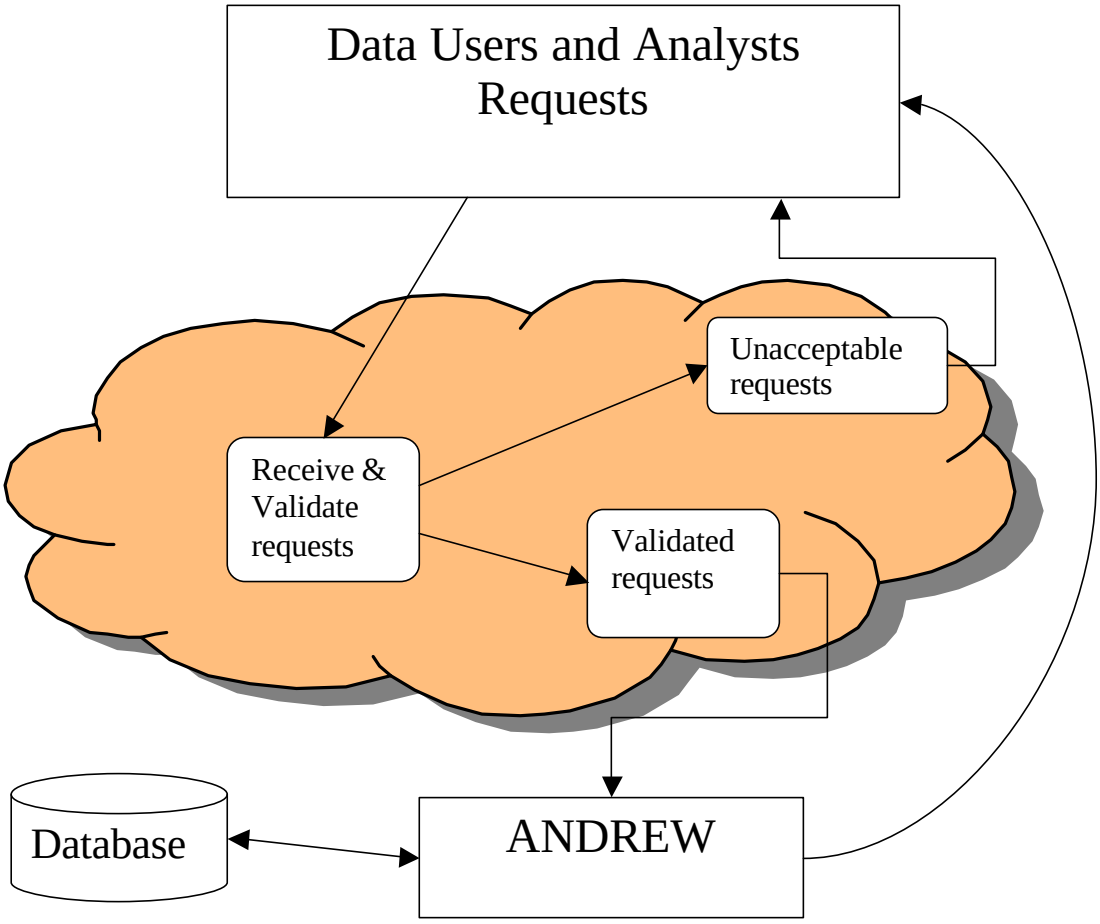
The following diagram is a graphic representation of the general schema and overall philosophy on which ANDREW is being developed. It shows various entities of the system and how they are related to each other.



4.2 The Web Component:

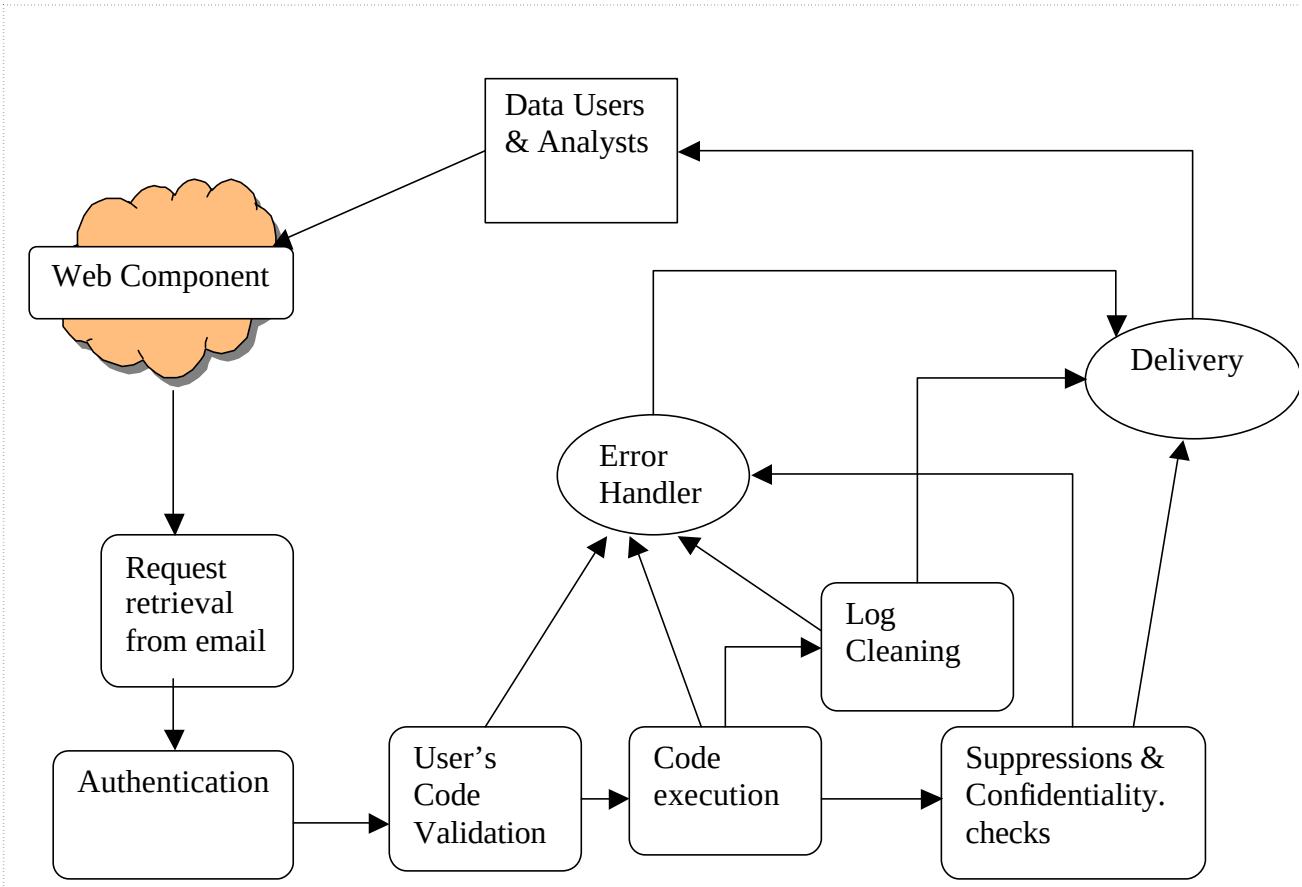
The following diagram depicts the user interface between the subscribers and the Web Component (WC) as well as the internal working mechanism of the WC. The WC receives data requests from the subscribers and determines

suitability/risk factor associated with the request based upon its analysis. The requests that pose no or manageable risks are passed on to the Main Component (MC). WC issues appropriate error messages to the requester if the risk is unacceptable.



4.3 The Main Component:

The following diagram shows internal functioning of the Main Component (MC). The request from the WC is authenticated and analyzed for the confidentiality risks. If the risk is manageable, the user code is executed. In case of any error detected during the processing of the request, the Error Handler formats an appropriate error message and arranges for its delivery. Eventually, if everything goes well, the output and log file (after it has been freed of any disclosure violations) are delivered to the subscriber via Email.



5 Summary:

Despite a number of inherent limitations and constraints, RDC's initial remote access system, ANDRE, has performed very well since its inception in 1998. However, with an ever growing recognition that new and improved technologies are needed, the RDC has undertaken to develop ANDREW, the next generation remote access system. When completed, ANDREW will represent a major advancement in the area of remote data access. Even so, it should not be considered the final product. It is a continuous software engineering process geared towards regular induction of improvements and enhancements to the system as dictated by the feedback from the user community and by the internal system performance analysis.