CMU-ITC-087
New Delhi, India '90

10th Internatinal Conference
on Computer Communication

Network Data Access

Mark Sherman, Mario Goertzel
Information Technology Center*
Carnegie Mellon University
4910 Forbes Avenue, Pittsburgh
PA 15213, USA
mss@andrew.cmu.edu, mg2p@andrew.cmu.edu

In this paper, we describe system requirements for allowing access to heterogeneous, preformatted data in a network environment. The system is based on providing an abstract description of data formats and data transformations in a way that data providers, data consumers and format converters may coexist tha cooperate on a network using a variety of hardware.

1 Background

The amount of machine-readable information being collected and distributed in all fields is staggering. For example, one can easily purchase information on construction codes, weather information, drug interactions, census data and stock market floor activities. New journals are devoted to describing newly available information sources on CD ROMs alone. These data are important to specialists in a variety of technical fields who, typically, are computer literate but are not programmers. This wealth of data is valueless until an expert in the appropriate field can bring the necessary analysis to bear.

Unfortunately, the data are provided on a medium and in a format convenient for the distributor. In some cases, a special program needs to be written to access the data. Thus, the importation of data is difficult, as it requires programmer expertise, which is expensive and time consuming. In many other cases, the information distributor or value-adder provides a self-contained, special-purpose access program for the data. However, this limits access to the data to the machine or program provided by the information source, limits sharing of data, and effectively prohibits simultaneous or interacting access to multiple data sources. Further, special purpose hardware, such as a video disk or multiple CD ROM readers may be required to examine the data, thereby limiting data access to a small set of properly configured machines.

2 Requirements

There is a clear need to provide a better way to let non-computer specialists import data. Part of our medical applications project is working on providing mechanisms for heterogeneous, multimedia data access in a network environment. Although we believe that we will handle simple text, we intend to handle a much richer variety of information. Each kind of data is described below.

* This work was performed as a joint project of Carnegie Mellon University and the IBM Corporation. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies of the IBM Corporation or Carnegie Mellon University.

International Conference on Computer Communication
ICCC 90, New Delhi, India

2.1 Structured Text

One of our data sources is an electronic medical library, containing books and periodicals. Some of these sources are marked up with SGML tags and some will be represented as ODA documents. We also expect some special purpose formats to be provided, both from documents produced on word processors that get introduced into the library and from typesetter tapes that publishers will be providing us. Some of the documents, such as those represented in ODA, will also contain formatting information.

2.2 Compressed Data

Many sources provide their information in a compressed format. Most of the time, the compression algorithm is published with the data, though some companies use proprietary compression algorithms. We need to accommodate both situations.

2.3 Graphic Data

Collections of clipart and raster images are becoming increasingly popular. Several widely used formats are available, including CGM, IGES, GKS and PostScript for structured graphics, and group 3, group 4 and TIFF for raster images. Medical image information uses a common, though proprietary, oval-array representation. We need to provide ways for converting between these formats.

2.4 Continuous-Time Data

An important collection of data for our purposes is continuous-time data. This includes audio, video, animation and process (or real time) data. There are three attributes of continuous-time data that are important. First, the data must meet more stringent delivery requirements than most data. Not only should the data arrive reliably, and in order, but it also must arrive at a certain rate. If the data arrive too quickly, there will be no buffer room and no time to retransmit. If the data arrive too slowly, gaps in the presentation of the data will affect its quality.

The second attribute is that data become stale. Examples of a continuous-time datum that can turn stale are the current temperature and a satellite photograph of a particular place. One should be able to introduce updates from these continuous-time data sources into a network environment as easily as a source frozen on a CD ROM.

The third attribute is a need for synchronization. Multiple streams of continuous-time data, such as video and audio, may require synchronization in their delivery and presentation. Therefore, part of the delivery requirements for retrieving continuous-time data would be their dependency on the delivery of other data.

2.5 Tupled Data

Many data collections are based on conventional database models using records as a storage model. Therefore, we need to provide a way to access a variety of database systems, including conventional database systems and newer multimedia, object-oriented database systems.

3 Related Work

Several groups have tried to provide access to services in a heterogeneous environments and to provide access to heterogeneous data. For example, the Mercury project at MIT [Liskov et al, 1988], the HCS project at the University of Washington and the RM system at the IBM Los Angeles Scientific Center all bring together services on a different kinds of machines. However, these groups have focused on providing a universal remote procedure call (RPC) facility. Although useful, an RPC mechanism is insufficient for our needs: it still is a programmer's level interface, it requires a priori knowledge of services for connection initiation and (with perhaps the exception of the Argon implementation in Mercury), does not provide a sophisticated way to describe data.

Several groups have tried to provide simultaneous access to diverse databases. For example the Multibase project at CCA, the Polypheme project at INRIA [Litwin, 1988], the Dataplex system at GM [Chung, 1990], the Linda system at Technical Research Centre of Finland [Wolski, 1989], and the Darwin system at Merrill Lynch [Rizzo and Strauss, 1988] attempt to provide a uniform data model to access various databases. The multi-model system developed at the Connecticut University [Demurjian and Hsiao, 1989] allows the user access to more than model at a time. These systems differ from our system in two important ways.

First, they did not consider multimedia or continuous-time media, whereas both are part of our requirements. Second, they were concerned with database updates propagating to all of the appropriate systems, whereas we are concentrating on data that is not changed by the user. Therefore, we have a much greater need for data description and a much lesser need for maintaining data consistency across data sources [Hsiao and Kamel, 1989].

4 Format Conversion

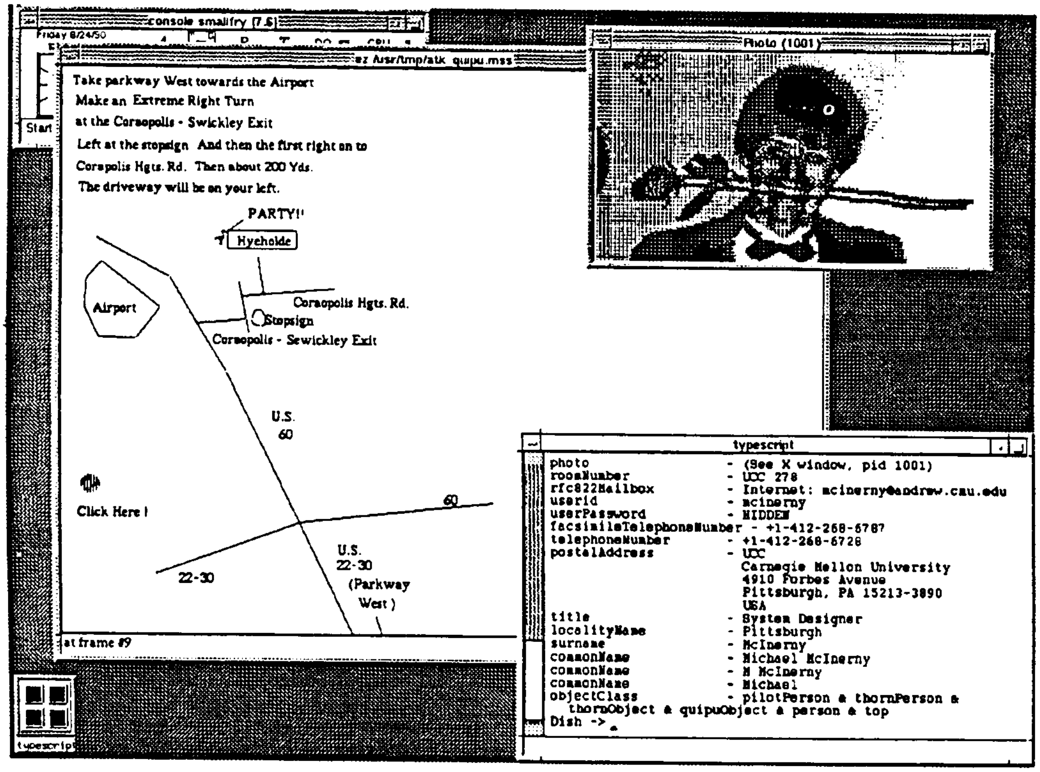
To accomplish format conversion, we intend to provide format servers on the network. These servers would advertise their operations through the use of x.500. For example, the PostScript Interpreter server would advertise itself as a format converter that could accept PostScript as input and produce a bitmap image as output. Clients needing bitmaps but finding PostScript would route PostScript data through that server in order to perform the needed transformations. We would have to expand the defined types (object identifiers) in x.500 for attribute values. We have already done this for our own multimedia format.

5 Strategy

One approach to providing uniform function in a heterogeneous environment is judicious use of standards. We have already taken the first step in providing support for ODA documents and conversions to other formats. As already indicated, we plan to use x.500 directory services for locating data sources and format converters. We operate primarily on TCP/IP networks, so our initial implementation is based on that transport, using the ISODE package as an intermediary between the OSI and TCP worlds.

6 Status

We have a conversion package that performs the transformation between a variety of raster formats. The converter system can be hooked into a network server. We have also expanded the capabilities of x.500 to allow the delivery of general multimedia information using our own multimedia format. Figure 1 shows the expanded system running with the dish utility provided by quipu package [Kille et al, 1990]: one window contains an image as provided by dish, while another shows an animation retrieved from the same entry. The animation was created



Take parkway West towards the Airport
Make an Extreme Right Turn
at the Corsopolis - Swickley Exit
Left at the stopsign And then the first right on to
Corapolis Hgts. Rd. Then about 200 Yds.
The driveway will be on your left.

```
photo                    - (See X window, pid 1001)
roomNumber               - UCC 278
rfc822Mailbox            - Internet: mcinerny@andrew.cmu.edu
userid                   - mcinerny
userPassword             - HIDDEN
facsimileTelephoneNumber - +1-412-268-6787
telephoneNumber          - +1-412-268-6728
postalAddress            - UCC
                           Carnegie Mellon University
                           4910 Forbes Avenue
                           Pittsburgh, PA 15213-3890
                           USA
title                    - System Designer
localityName             - Pittsburgh
surname                  - McInerny
commonName               - Michael McInerny
commonName               - M McInerny
commonName               - Michael
objectClass              - pilotPerson & thornPerson &
   thornObject & quipuObject & person & top
Dish ->
```

Figure 1: x.500 integrated with Multimedia Attributes

by our local multimedia system, while the image was created by filtering our local image formats into those acceptable to quipu and dish. The same filter would be basis of the network format converter.

7 Summary

There is a great need for communities of people to import preformatted data and share them in a network environment. Our project attempts to ameliorate this problem by providing a high level description of data formats and of format converters that can be accessed across a network using machine-independent protocols.

8 References

[Chung, 1990]
Chung. Chin-Wan, DATAPLEX: An Access to Heterogeneous Distributed Databases, *Communications of the ACM*, Vol. 33, No. 1, January 1990, p. 70-80.

[Demurjian and Hsiao, 1989]
Demurjian. S.A. and Hsiao. D.K., The multi-model database system, *Conference Proceedings of the Eighth Annual International Phoenix Conference on Computers and Communications*, IEEE Computer Society Press, March 22-24, 1989, p. 439-45.

[Hsiao and Kamel, 1989]
Hsiao. D.K. and Kamel. M.N., Heterogeneous databases: proliferations, issues, and solutions, *IEEE Transactions on Knowledge and Data Engineering*, Vol.1, No.1, March 1989, p. 45-62.

[Kille et al, 1990]
Kille. Stephen E., Robbins. Colin J., Roe. Michael, Turland. Alan, *The ISO Development Environment: User's Manual, Vol. 5: QUIPU*, SPI Inc., 420 Whisman Court, Mountain View, CA., January 12, 1990.

[Liskov et al, 1988]
Liskov. Barbara, Bloom. Toby, Gifford. David, Scheifler. Robert, and Weihl. William, Heterogeneous Computing: A High-Level Communication Mechanism, *Chaos into Order: Proceedings of CIPS Edmonton '88*, Canadian Information Processing Society, Edmonton Section, Edmonton, Alberta, Canada, p 214- 222.

[Litwin, 1988]
Litwin. W., From database systems to multidatabase systems: why and how, *Proceedings of the Sixth British National Conference on Databases*, Cambridge University Press , July 1988, p. 61-88.

[Rizzo and Strauss, 1988]
Rizzo. Tony, Strauss. Karen, "DARWIN: Merrill Lynch Develops a New Workstation Based on Windows 2.03," *Microsoft Systems Journal*, Vol. 3, No. 4, July 198, p. 1-12.

[Wolski, 1989]
Wolski. A., LINDA: a system for loosely integrated databases, *Proceedings Fifth International Conference on Data Engineering*, IEEE Computer Society Press, Feb. 6-10, 1989, p. 66-73.

Mark Sherman is a scientist at the Information Technology Center at Carnegie Mellon University. He has a PhD from CMU in Computer Science. His current research interests include distributed systems and user interface development systems.

Mario Goertzel is a system programmer at the Information Technology Center at Carnegie Mellon University. He is working towards his degree in mathematics and computer science at CMU. His primary interests are in heterogeneous systems.