USE OF HETEROGENEOUS DATA SOURCES: THREE CASE STUDIES

by

DAVID BRADLEY GODES

B.S., Finance and International Business
University of Pennsylvania
(1985)

Submitted to the M.I.T. Sloan School of Management
in Partial Fulfillment of
the Requirements of the Degree of
Master of Science in Management

at the

Massachusetts Institute of Technology

May 1989

Signature of
Author_____
M.I.T. Sloan School of Management
May 22, 1989

Certified
by_____
Professor Stuart Madnick
Professor of Management Science
Thesis Supervisor

Accepted
by_____
Jeffrey A. Barks
Associate Dean, Master's and Bachelor's Programs

USE OF HETEROGENEOUS DATA SOURCES: THREE CASE STUDIES

by

DAVID BRADLEY GODES

Submitted to the M.I.T. Sloan School of Management
on May 22, 1989 in Partial Fulfillment of
the Requirements of the Degree of
Master of Science in Management

ABSTRACT

This Paper discusses the process of financial analysis within the context of Composite Information Systems (CIS) through an analysis of three cases. This was written in conjunction with the Composite Information Systems/Tool Kit (CIS/TK) research project at the M.I.T. International Financial Research Center. A primary purpose of the paper was to identify, document, and understand the needs and problems of users of Composite Information Systems. The analysis makes use of the delineation between "physical connectivity" and "logical connectivity".

The first case study is from the academic domain. It is an event study of the potential differential effects of the October, 1988 stock market crash across a sample group of companies. The second case study involves CitiCorp's North American Investment Bank (NAIB) and their attempt to integrate more than twenty different processing systems. Their task is made even more difficult by the fact that there are not one but three main groups demanding this integration, each with a somewhat different goal. Finally, the third case study, also from CitiCorp, involves the Corporate Financial Analyst Department (CFAD) in the institutional bank. They make use of many different types of data and the paper investigates the problems that they face in integrating the data on both an inter- and an intra-database level.

Each of the three case studies takes the following form: a description of the "problem", outline of the nature of the data involved, and documentation of the problems that one would face in integration. Finally, these problems are related back to the CIS/TK project.

Thesis Supervisor: Dr. Stuart Madnick
Title: Professor of Management Science

Acknowledgements

I would like to thank everybody who made sure that this thesis was at once a learning experience, a challenging and entertaining project, and the successful "partial fulfillment" of my academic requirements:

CitiCorp: It would have been impossible to complete this without their willingness to contribute so greatly. In fact, the only thing more difficult would be to list all of them that have helped, but here's a try: Judy Pessin, Dorothy Conroy, John Remmert, Bud Berro, Evan Picoult, Gustavo Schwartz, Dan Schutzer, Helga Oser, David Moore, Steve Ellis, George Levitt, Gary Geresi, Mary Cirillo, Nestor Hernandez, Ken Wormser, Tracey Peter, and David Lipfert. Thank to all!

Stu Madnick: For his direction and particularly for his ability to push me several times beyond where I thought this project might have ended.

Mia Paget, Bertrand Rigaldies, and Bob Goldberg: For their invaluable insight into both CIS/TK's technical and conceptual issues.

Jodi: The woman of my dreams: past, present and future. For putting up with a poor, irritable, unorganized, procrastinating student for the last two months and still not asking what CIS/TK stood for.

Murray: Rope, Jim, Bob, Bob, J.C., Burly, Gulley, and the MBAs Ash and Morty.

Table of Contents

CHAPTER 1: TECHNOLOGY (AND DATA) IN THE FINANCIAL SERVICES INDUSTRY

The "information technology revolution" has become such a great part of our lives, both at home and at work, that we seldom take the time to sit back and truly understand exactly what it has done for (or to) us. One might argue that the "moving target" nature of the related technologies (and the speed with which they move) precludes such an evaluatory analysis. In this first chapter, I will attempt to briefly describe the role of information technology in the corporation in general and then specifically in the financial services sector. The discussion of the financial services industry's use of IT will be preceded by a brief outline of the structural changes that have taken place within the industry in the past decade. This outline will serve as background for understanding the changing role of technology in the industry. Finally, I will put forth a general description of the nature of <u>data</u> in the corporation, its characteristics and importance. This in turn will help us understand the very real difficulty of integrating heterogeneous computer systems.

Following this chapter, the reader should have a good understanding of why a company may want to develop a Composite Information System (CIS), what they should consider when making the technology decision, and most importantly why the implementation of such a system is so difficult. The reader will also be familiarized with a popular example of the successful implementation of IT: the Financial Services Industry. This industry will serve as the context for the examples of such integration in future chapters.

I. Technology and the Corporation

The literature of the past decade has contained a great deal of discussion about the relationship between the corporation and Information Technology (IT). Such popular topics include the structural changes that technology has catalyzed in specific industries (Parsons [1983]), the way in which strategy and technology are related (Porter[1979]), and the ways in which a company might integrate the planning systems of strategy and technology(Henderson and Sifonis [1988], Henderson, Rockart and Sifonis[1984]). I will highlight three of the major contributions that are relevant to this analysis of Composite Information Systems. They discuss in turn: (1) How IT, if used correctly, can help the user to gain a strategic advantage in the specific industry; (2) The different stages, or "eras", through which corporations' usage of IT has moved, and continues to move; and (3) The variables that one must contend with when implementing a strategy based on, or simply including, IT.

A. What Can IT Do For Me?

Michael Porter and Victor Millar have contributed an important piece to the usage of information technology to achieve a competitive

advantage[1] . In this, they outline the three general ways in which IT has altered the playing field in many industries:

• **Changing Industry Structure**: It is clear that in many cases IT has changed the "rules of the game." I assume a familiarity on behalf of the reader with Michael Porter's industry structure framework which identifies the five forces that comprise the structure as: Buyers, Sellers, Substitutes, New Entrants, and Internal Rivalry. The relative "power" of these constituencies essentially defines the "structure." Therefore anything that changes this relative power has the potential to change the industry structure and therefore the relative profitability of the players among, and within, the constituencies. IT is doing just that. The medical products distributor American Hospital Supply (now part of Baxter Travenol), for example, has enhanced its power within the its industry by creating an inter-corporate CIS which ties them directly to the order processing system of their customers. This has since become an essential piece of the marketing strategy for any player to achieve any success in that industry.

• **Creating Competitive Advantage**: According to Porter, there are two generic ways to gain advantage: low cost and differentiation. So, again, to the extent that a company's use of IT aids them in the pursuit of either of these ends, the successful usage of IT may confer on that company a strategic advantage.

---

[1] Porter, Michael and Millar, Victor A., "How Information Gives you Competitive Advantage," Harvard Business Review, July-August 1985.

Further, Porter and Millar also point out that one more way in which IT might aid in the creation of such an advantage would be in the broadening of a company's "competitive scope". This might include geographic as well as business scope. A good example is the ability that USA Today developed in offering a truly national newspaper. Without IT, this would never have been possible.

• **Spawning New Businesses:** Finally, IT might allow a company to get into an entirely different business by helping them to leverage a certain strength (or overcome a weakness) that they possess. Such leveragable strengths might include a loyal customer base (which prompted Sears' diversification into financial services) or excess data processing capacity (Eastman-Kodak has entered the long-distance phone service business by offering service through its internal telecommunications network to external customers). Merrill Lynch's launching of the Cash Management Account (CMA) which combined three distinct financial products into one would never have been possible had the technology not been able to provide the level of integration necessary between the processing systems for each product.

So, it should be clear that, for many companies, IT is playing an extremely large role in their "value chains"[2] . Further, Composite Information Systems are becoming an important tool in the corporate strategist's toolbox. The examples given should serve as a reminder, to which I will constantly return throughout this thesis, that there is, in

---

[2] More on the Value Chain concept can be found in Porter, Michael, Competitive Strategy, New York: Free Press, 1980.

fact a reason for companies to be deeply interested in connectivity and the related technologies: if used properly, one can gain a substantial and perhaps sustainable strategic advantage.

B. The "Eras" of Technology

As alluded to in the opening paragraph, technology is a moving target. The ways in which we interacted with technology in the 1970's are very different from the ways in which we interact today (and from the ways in which we will interact in the 90's). Of course, different organizations are affected by the changing technologies at different paces. The factors which might influence the pace at which a company takes advantage of changing technologies include: the company's size, the technology intensity of their value chain, the stages of the life cycle in which their products exist, and the age of their current technology (which dictates to some extent when they will be "in the market" for technology again). A very useful paradigm for understanding the ways in which a corporation might use IT, and how this usage might change over time, is offered by Jack Rockart in his delineation of the "four eras of information technology"[3]. The first three of these "Eras" are outlined in Fig. 1-1 and they all can be summarized as follows:

> **Era 1**: This is the "accounting era" in which the main use of information technology is for the processing of very data-intensive, accounting-related applications such as general ledgers,

[3] Rockart, John G., and Bullen, Christine V. eds, The Rise of Managerial Computing, Illinois: Dow-Jones-Irwin, 1986, Introduction.

Era		Applications	Hardware	I/S Tools	I/S Mgmt
(1)	Accounting	Payroll A/P General Ledger	Centralized Maxis Batch	Project Mgmt Capacity Mgmt COBOL	Line
(2)	Operations	Order-Entry	Decentral-ized Minis On-Line	Project Mgmt Capacity Mgmt COBOL	Matrix
(3)	Information	DSS XSS	Cent/Decent On-Line	Evolution-ary Dsgn. Budget RAMIS FOCUS	Matrix Staff Support Everyone a user

Figure 1-1: The Four Eras of IT by Jack Rockart

accounts receivable, etc. It is generally characterized by batch systems which are run by a DP organization that resides below the financial organization in the corporate structure.

Era 2: This is the era which is characterized by "operational systems". That is, the computer has left the accountant's office and begins to aid in other areas such as order processing, sales tracking, production data-gathering, etc. Again, batch systems tend to predominate, and the CFO or Controller tends to retain control over the bulk of the IT resources.

Era 3: This era is characterized by a greater integration of strategy and technology with a resulting emphasis on more IT planning within the organization. We also begin to see more on-line systems in the organization as well as the proliferation of independent "data centers" as IT groups separate from the financial people and establish their independence. Further, the evolution of the DSS (Decision Support System) has begun as well as its upscale cousin, the EIS (Executive Support System). Here, computers begin to aid in the process of problem solving rather than simply providing data as an input to that process.

Era 4: Rockart refers to this as the era of a "wired society" where there exist multiple levels of connectivity: inter-corporate, intra-corporate, etc. There also is a tendency toward a senior staff-level IT "guru" sometimes known as a Chief Information Officer (CIO) serving to bridge the gap between the IT people and the line. Significantly, in this era the line itself tends to take more of a lead

> in pushing IT developments and implementation of these developments. IT, as an organizational entity, thus reacts to technology-based ideas by the line rather than proactively pushing new developments the other way.

Again, an "era" in this respect is not characterized by conventional, standard temporal markings such as years or decades, but is organization-specific and depends on the factors described above (size, etc.).

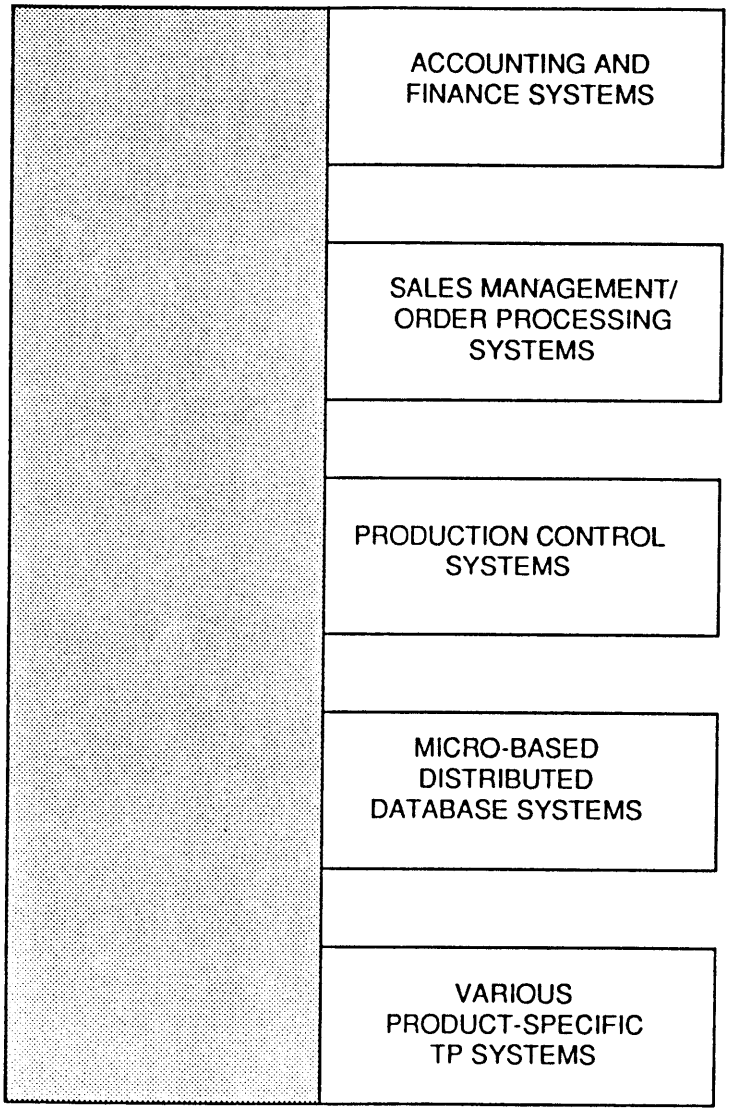
In the context of Composite Information Systems, then, Rockart's chronology gives us a fairly good idea as to both the importance of their effective implementation and the difficulty in doing so. Their implementation is vital simply because many, if not most, companies are recognizing the advantages of a more "wired" IT environment. The examples of American Hospital Supply and Merrill Lynch and many others have shown other companies the value that can be added through CIS innovation (or innovative use of current technologies). Further, Rockart demonstrates one of the reasons why the integration itself is so difficult as it implicitly describes the evolutionary nature of any single company's use of IT. In the simplest case, first the accountants used IT. Then, the sales department found that their order processing could be handled efficiently on a different , stand-alone system. Then manufacturing and marketing decided to develop stand-alone applications including microcomputer-based databases and decision support tools ranging from the shop floor to the CEO's office. This might be described as the development of "stovepipe" systems (the derivation of their name is made clear in Fig. 1-2) during eras 2 and 3.

This might have been fine, however, as long as the company's needs were satisfied within the Era 3 environment. That is, it was fine until we felt a need for our manufacturing systems to "talk to" our marketing systems (for example, so the salespeople could update plant managers on the status of their custom orders). It is with this Era 3 infrastructure (stovepipe systems) with which many companies are entering Era 4 and hoping to reap the benefits of the "wired society". The difficulty in doing so is manifest.

The natural corollary would imply that those beginning their use of IT in Era 4 will have a very easy time. Easier, perhaps, but as the next section points out, there is a great deal more to consider than simply technology, and therefore the road to a wired environment and CIS may still not be a completely smooth one.

C. How Can a Company Make IT Work?

Porter and Millar have clearly shown the advantages of IT or <u>why</u> a company might want to use a technology such as a CIS. Rockart has shown the evolution of <u>how</u> companies have used these technologies for the creation of advantage as well for the support of their ongoing business. This final literature review section will discuss Rockart and Scott-Morton's description of the other (non-technology) variables that one must consider when attempting to implement technological


ACCOUNTING AND FINANCE SYSTEMS
SALES MANAGEMENT/ ORDER PROCESSING SYSTEMS
PRODUCTION CONTROL SYSTEMS
MICRO-BASED DISTRIBUTED DATABASE SYSTEMS
VARIOUS PRODUCT-SPECIFIC TP SYSTEMS

solutions to strategic or operational difficulties[4]. Fig. 1-3 shows these variables and their interrelationships. Other than the sheer complexity of the network (and therefore of the decision itself), it is important to note as well the lack of a beginning or an end (or any true directionality) in this figure. That is, there is no hard and fast rule as to the causal ordering of these factors. Several people have discussed this phenomenon with respect to two of the major components: technology and strategy. They have made it clear that technology can both be driven by a company's strategy (such as in the case of an investment in the data processing technologies by Merrill Lynch to support their extremely successful Cash Management Account) and be a driver of strategy (such as in the case of Sears, whose large computing resources and resulting customer database opened up to them a brand new business: financial services).

Expanding this concept to the larger domain of the other variables, then, we can say that they all are related but none follows directly from any other as a rule. A change to any one of them could easily trigger a "misalignment" in one or many of the others. For example, a change in the "roles and responsibilities" might conceivably lead to stress on the current "organizational structure". This might then cause a change in the necessary technological resources which support that part of the organization, which might easily lead to change in any of the other factors. So, it is important to understand the nature of these

---

[4] Rockart, John G., Scott-Morton, Michael S., "Information Technology, Integration, and Organizational Change," MIT Management in the 1990's Working Paper #86-017, 1986.

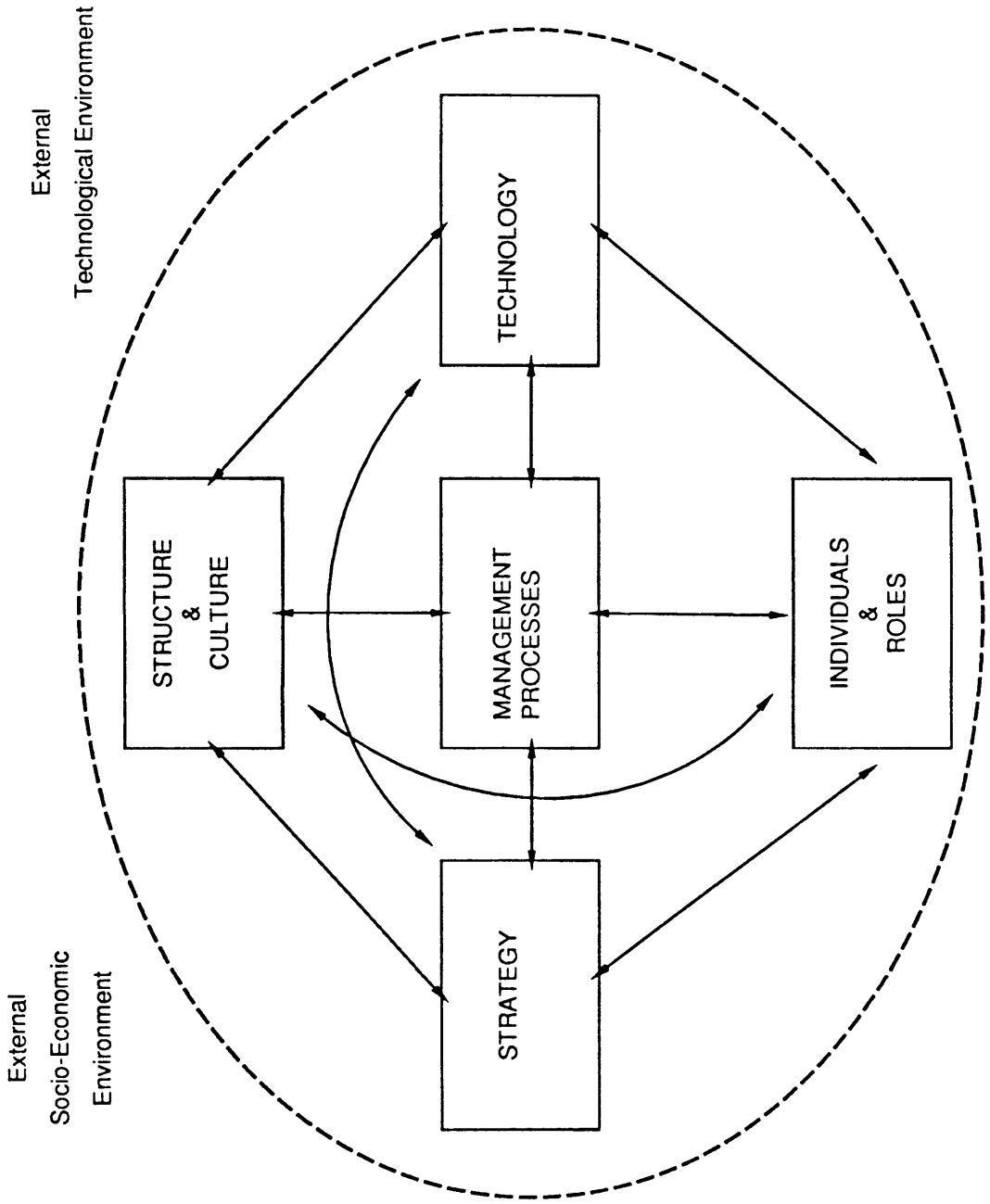


Fig. 1-3: The Management in the 1990's Framework

interrelationships within the organization and the constraints that they put on technology-related decisions.

Combining this paradigm with the previous two, we get a fairly good picture of many of the "big" issues surrounding Composite Information Systems, the main topic of this thesis. That is, many companies see a need (or desire) to link their systems in some way. However, for evolutionary reasons these systems were often not meant to work together. This might take the form of a different platform, different standards, operating systems, formats, etc. Further, even were a company to somehow ensure the compatibility of these factors beforehand, there would be a host of other issues that would need to be balanced when making the technology decision. This answers the very real question posed by companies that may ignore (or have avoided) the constraints of the previous eras: "Why not just build one big system?". While this may make sense technologically (and even strategically in some cases), the other factors in the Rockart/Scott-Morton framework generally make sure that this is not a viable option. In some sense, they are often "stuck with" multiple systems in a distributed processing environment (due to the constraints of these "other" factors in the Rockart/Scott-Morton model) and to reap the benefits of integration across these heterogeneous systems they must clear the many hurdles that are discussed, described, and analyzed in the remainder of this thesis.

II. The Financial Services Industry[5]

The early 1980's saw a torrent of deregulatory legislation in a wide range of industries including transportation, telecommunications and financial services. The combination of industry lobbying, a free-market administration, and sound economics led to the lowering of decades-old restrictions on the ways these and other industries do business. The aftereffects were, and continue to be, profound. In transportation, we saw the birth (and essentially the death) of a new breed of competitor: the discount airline. In telecommunications, a similar outcome has evolved with first many and now a few lower-cost long distance carriers. The changes to the Financial Services Industry (FSI), while just as profound, seem to have occurred over a longer time frame. It seems that it has been more a case of "creeping deregulation" than the equivalent of the breakup of AT&T and the dismantling of the fare and route structures which supported airline regulation.

A. The Changing Face of the FSI

While I separate for clarity the section on the changing structure of the FSI and that on the role of technology in the industry, I would like to make it perfectly clear at the outset that they are inextricably intertwined. Not only has the deregulation led to new product offerings

---

[5] Much of the information for this section has come from "The Evolving Financial Services Industry," HBS Case #9-183-007, Harvard Business School, Boston, MA, 1983.

which in turn has led to new IT applications, but the technological infrastructure itself has also allowed some of the players to profitably enter new markets. It has clearly been a two-way street.

Pre-deregulation, the FSI would have best been described as an "institutionally-based" industry. This refers to the tendency at that time for the industry to be divided into segments defined by the institutions themselves. That is, there was the insurance industry, the commercial banking industry, the investment banking industry, the brokerage industry, etc. And traditionally, the players, as defined, stayed within their segments.

The explicit regulations that had existed for years and preserved this structure essentially took four general forms:

- **Geographic**: This restricted the diversification of certain institutions into other geographic markets. A perfect example of this is the (rapidly eroding) regulation against interstate commercial banking.

- **Product Line**: This restricted the products which any player could offer. The Glass-Steagall Act is an example. This draws a line between the commercial banking and the securities businesses. However, this is also becoming obsolete as many players have attempted and succeeded in ventures "across the line" which even ten years ago may have led to a call for swift action by the SEC and other various regulatory bodies.

- **Pricing**: These restrictions, still present in most of the institutional segments of the FSI, restrict the pricing strategies of the FSI players. They range from APR disclosure requirements in consumer lending to limits on the on the investment returns offered by "whole life" insurance policies.

- **Entry**: Finally, entry into each of the institutional segments was restricted by a number of explicit regulations including asset size and capitalization requirements.

Not surprisingly, a by-product of this institutional mindset and other factors (such as a generally accepted opinion that too much competition would harm the end consumer) was that the level of competition in the industry was, by most accounts, not as high as that of most other, more free-market-based, industries. Further, there were few economies of scale to exploit in this industry and small players therefore found it relatively easy (if the entry issue was overcome) to find a niche and compete successfully.

However, this changed in the late 70's and early 80's. A combination of deregulatory legislation, changing economics (particularly in the form of extremely high inflation rates, which led investors to look for investments which would protect their returns in such an inflationary environment, and away from long-term low-return investments such as certain forms of life insurance), and technology led to the gradual erosion of this institutional mindset and the evolution of a more market-based industry. In this new structure, the competitors tend to face a more "market-based" segmentation, organizing around market

segments rather than product groups. Within this environment, to compete successfully a player often must offer a full range of services (many crossing the "lines" that had been drawn in the past) to a market segment or segments.

B. Technology and the Evolving FSI

Again, the technological changes in the FSI cannot be considered as simply a by-product of the changes in the industry. One must consider also the effect that technology itself had on the industry.

The FSI, more than just about any industry, has been forever linked to technology. Among the first users of computers, the FSI's growth would never have been as rapid had it not been for the growth in processing, monitoring, and storing capacity that was made possible beginning in the 1950's with the computer revolution. This is true of every segment of the industry. A good indicator of at least the commercial banking segment's dependence on computers is the fact that to process the more than 40 billion checks written annually in the United States, commercial banks would require the services of over half of the U.S. workforce! So, it is not surprising that the futures of IT and FSI have been be closely linked.

Referring back to Michael Porter, there have been countless examples in the FSI of competitors using IT to achieve advantage in several ways: Cost Advantage was the initial driver of CitiCorp's proliferation of Automatic Teller Machines (ATM's); Differentiation was achieved by

Merrill Lynch when they first introduced their Cash Management Account (CMA). The FSI is rich with such technology-based strategic positioning.

Beyond the impact of IT on company-level strategies, it is crucial to understand the impact that IT has had on the FSI as a whole. First, it has changed the nature of the restriction on geographic scope. For example, the evolution of national (and international) ATM networks has reduced the need to "be everywhere". Other such examples include the growth of "electronic banking", as well as the development of an electronic stock market in New York and (moreso) in London.

Further, the increasing intensity of IT all along the value chains of the major FSI players has increased the potential for economies of scale in the industry. This has thereby further improved the chances of the larger players (cetis paribus) to succeed in the changing industry. Now, "being big" means that a player might be able to build their customer base using IT (for example through an ATM network) and then leverage this asset into other product lines. It is becoming increasingly probable that the FSI will be dominated by several full-service giants- CitiCorp, American Express, Prudential, etc.- while smaller players will find it even more difficult to compete against the economies of these giants.

To summarize, the Financial Services Industry continues to change today. The role of technology is multi-fold. IT serves both as a strategic weapon and as a constraint on strategic thrusts, depending on which technology and who is using it. However, one point is extremely clear: to succeed in the industry, a company must understand IT and take advantage of what it has to offer. As will be discussed later in the chapters on CitiCorp, to take advantage of IT is far more than simply a question of technology, but one of balancing the concerns of organization, technology, and strategy.

III. The Character of Data

This section will put forth some general propositions about "data". It will include a definition of the term "data"; a discussion of the uses, and processing, of data; and a brief review of the nature and characteristics of data. The goal of this brief discussion is to familiarize the reader with the concepts of data integration, data interfaces, and data processing to which I will refer throughout the rest of this thesis.

A. What is "Data"

A "datum" (the singular form of "data") is described by Webster's dictionary as "**1.** Something given or admitted, as a fact on which an inference is based. **2.** Something, actual or assumed, used as a basis for reckoning." These definitions are extremely interesting in this context for several reasons. First, they are clearly non-technical. For those who think that "data processing" arose in the 1950's with the ENIAC, this may come as some surprise. The fact is that the computer has certainly allowed us to increase manyfold our capacity for DP, but we had been doing it ourselves since time immemorial. In fact, a great deal of data manipulation and processing is still performed manually (and mentally). A key question raised by this thesis is to what extent do we want to remove the human element from the loop?

Second, data is not data unless it is to be used for the attempted derivation of the solution to a problem, the answer to a question, or to be used as a small component toward these ends. We, whether in a

commercial or personal context, use data to solve problems. The process through which we do so can be defined fairly generally by the model in Fig. 1-4. This notes the various stages of the problem solving process. Of course, depending on the problem, this might a be more or less recursive process. However, the general order of these stages should be generally standard across problems.

This paradigm could easily be used to model many of the data intensive operations of a commercial bank. For example, the monitoring of a bank's risk exposure, a function I will discuss later at CitiCorp, consists of gathering data about the various securities and positions that the bank owns (the "definition of the problem" stage occurs once and is likely only updated at odd intervals), combining that data into groups of similar (and perhaps offsetting) sensitivity groups, and analyzing that data to evaluate the limits that are currently set to constrain the activities of traders. Finally, the "answer" might come in the form of new limits, or a confirmation that the current limits are "okay". I will discuss this process in more depth later as well as the risk management function's usage of heterogeneous data sources. Simply, the necessary data was defined, gathered, processed (combined), and analyzed in search of an answer to a problem.

Given this model, it should be clear that the integration of the data comes about in the second and third stages. In fact, it is likely that integration is a key component of the functionality of these two stages. Further, it should be clear that the responsibility for performing any of the five stages could be given to either a human or a machine, depending on the level of complexity of the problem. It then becomes a

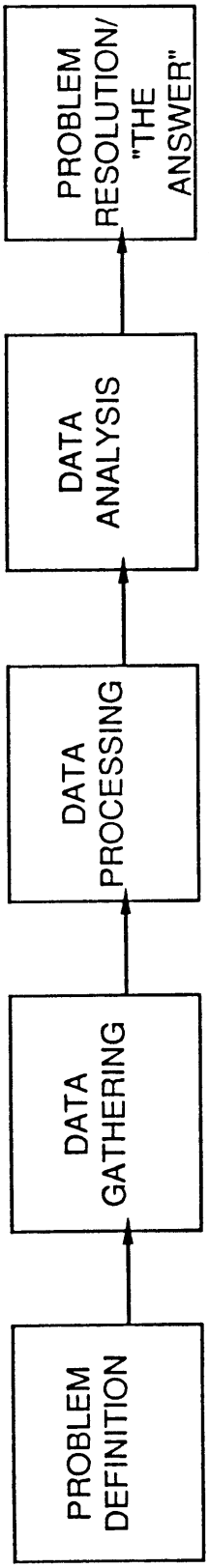


Fig. 1-4: The Generic Problem-Solving Process

question of which would be the most effective and efficient performer of these tasks, man or machine? Why even attempt to automate the human process? The answer is certainly case-specific. Tasks that require repetitive, high-speed computations are likely best handled by a machine, but there is clearly a grey line. The ultimate objective of the implementation of technological replacement is clearly to free up the human to perform those tasks that he/she is best doing: those requiring reasoning. This is what CIS/TK is designed to do through its role in aiding the integration of heterogeneous data as explained on the next chapter.

B. The Many Faces of Data

Data, like any object in our universe, has a multitude of characteristics (or properties, to use computer jargon). It will be very helpful later on in the discussion of data integration to understand exactly what some of these characteristics are. Once we understand these features of data, we will have understood where data can differ, and therefore where there must be some level of intelligent integration to be able to use them together in the fourth stage of the process ("analysis"). The main characteristics that I would like to highlight are:

- **Location**: This characteristic describes where the specific data group resides. Common locations might be the corporate mainframe, the distributed PC-based databases, on-line

information services, newspapers, the grey matter of an analyst, etc.

• **Meaning**: By this, I refer to meaning at its lowest level. For example, the field in the fourth column of a database might "mean" or represent the deutschemark sales of the foreign exchange group on Tuesday or it might "mean" the deutschemark sales of a specific foreign exchange salesperson on Friday. I call this "low-level" because the sales figure may also "mean" that a limit was exceeded in the foreign exchange trading department or "mean" that the salesperson performed phenomenally. However, this meaning is probably better considered as an output from some stage of processing and/or analysis. It is this understanding of the underlying concept (the "meaning") which the data is representing that allows the user/analyst to actually gain value through the use of the data.

• **Value**: This is obviously the "level" of that concept which the datum is representing. A '4' in the field mentioned above could thus signify that the trading department bought 4 million deutschemarks or that the salesperson sold 4 thousand deutschemarks. This all depends on the next characteristic, "format".

• **Format**: This describes the way in which the data is represented in its "location" (as opposed to on a report, or on a screen). The choice of format takes into account such factors as the necessary precision of the data, the typical orders of

magnitude, as well as system or database storage requirements. An example would be the 4 million deutschemarks mentioned above. This might be represented as a '4', a '4000', a '4000000', a '4,000,000.00', etc. Further, and this clearly overlaps with the "meaning" characteristic above, one must understand that the 4 million figure is in deutschemarks. It is unlikely that the format would impinge on the analysis or processing of the data. However, it is clearly essential for useful processing and analysis that the format be known to the user/integrator

Source: While in some cases (particularly with on-line information services) it is likely that this characteristic and that of location above would be one in the same, it isn't the case with all such data. In some applications, such as in financial analysis, the credibility of the source plays a major role in the analysis of the data, and thus the identity of the source must be known in these cases.

Of course, one could list many other characteristics of data as well as further split (or combine) the characteristics that I have mentioned above. However, for this analysis it should suffice to have this general understanding of the nature of data, and the ways in which various data sets may differ. It then brings us to the main problem at hand: integrating heterogeneous data into a single analysis.

C. The Data Interface

This final section of the introduction will set the scene for the chapters to come which will describe the actual difficulties that people have had and are having in integrating data. Here I will describe the concept of a "data interface" which will be a main topic throughout the thesis.

Given that a problem, as defined by the user, requires the use of heterogeneous data in an integrated manner, it immediately becomes clear that there must be an understanding, and a reconciliation, of all of the major characteristics outlined above as the data is brought together. This reconciliation will take place during the data gathering and processing stages, depending on the specific data characteristic as well as on the integration strategy of the users:

Location: This must be the same across all of the data sets. This might involve the batch downloading of data from several mainframes into Lotus 1-2-3 spreadsheets to be merged into one spreadsheet for analysis as well as the on-line access of stock prices which are integrated with other data for generation of "buy/sell" recommendations by brokers. It might also involve an analyst reading a newspaper on the subway (thereby changing the "location" of certain of those reported data from the page to his/her head) and integrating this with internally-produced databases at the office about the same subject. Either way, it is clear the data must eventually reside, in some form, at the same location to be used in an integrated analysis.

Meaning: To the extent that there are many meanings (i.e. data definitions) to remember in any substantial database, the

integration of many databases complicates the data processing function greatly. Further complicating matters is the fact that there likely exist data with the same meaning in multiple databases. This means that this commonality must be recognized (which may be difficult due to the likelihood of different naming conventions) as well as the fact that there must exist a process for resolving contradictions between data with the same meaning but different value and/or format.

For example, one database may have company-level stock information and another may have company-level bond information. They might be used together to generate a company valuation. One example of data with the same meaning would be the company name (representing the same concept: the identification of the corporation for whom the data is reported). Therefore, when integrating the databases, this fact must be recognized and to use the data effectively, we must resolve inconsistencies between the values (i.e. different company names) and formats (i.e. different representations of the company's name).

Value: As described in the previous section, it is essential to resolve contradictions involving data with the same meaning but different value.

Format: Again, this must be understood and a common formatting and scaling strategy, which is driven by the ultimate use of the data, must be devised and the databases must be

converted into this format. An example might be the integration of data across several sales branches for a computer company. The West Coast branch might keep their sales in their local database formatted in thousands of dollars. However, the Peruvian branch, which may not have been doing as well, might record their sales in Sols (not thousands of sols). For the Swedish parent company to analyze worldwide sales en toto, they must recognize these differences and standardize the formats into one scale and one currency. As will be discussed in more depth later, this standardization can be done in one or several of many locations.

Source: Again, this identity must be preserved in some specific applications.

This reconciliation process is described graphically in Fig. 1-5. The locus of the union (but not necessarily the standardization, as will be discussed in Chapter 5) of these heterogeneous data occurs at what I will refer to as the "Data Interface". Clearly, it takes a certain level of intelligence, both specific and general, to perform such an integration. It should also be clear that the interface could again be either human or non-human. Currently, it is safe to say that much of the integration of data is done by human data interfaces. A good example of this is given in Frank, Madnick and Wang [1987] where they describe an international commercial bank's nightly manual "tape hand offs" which integrate heterogeneous data for analysis.

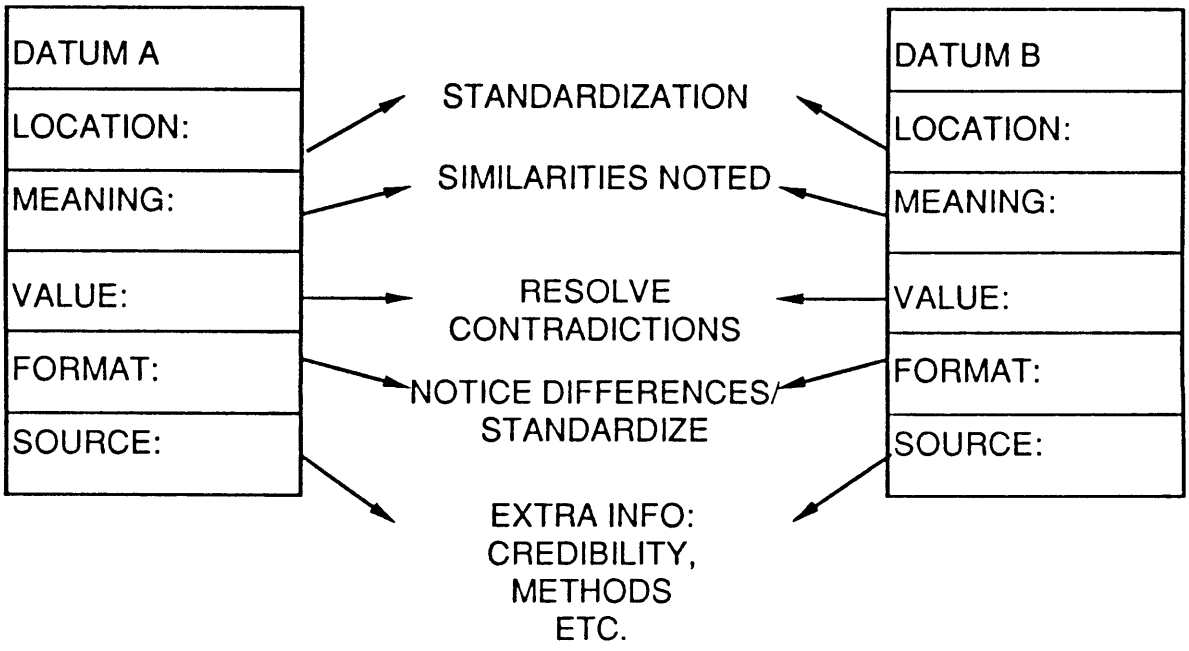


Fig. 1-5: Reconciliation at the Data Interface

The reasons for the predomination of human interfaces range greatly. However, the main reason is likely very simple: this is an extremely difficult and knowledge-intensive function to perform. The CIS/TK project (Composite Information Systems/Tool Kit) at MIT, a major multi-disciplinary research project being led by Professor Stuart Madnick is aimed at addressing this difficult issue and at providing users of heterogeneous databases and Composite Information Systems with the ability to perform some portion of this integration using a non-human technical data interface. The next chapter will describe the CIS/TK project in depth.

CHAPTER 2: THE COMPOSITE INFORMATION SYSTEMS/TOOL KIT (CIS/TK) RESEARCH PROJECT

Chapter 1 provided many different types of background information in an effort to familiarize the reader with the general nature of the problem of integrating heterogeneous data sources in the Financial Services Industry. This chapter will now go into more depth on this specific topic and will highlight the research being performed at the M.I.T. Sloan School of Management under the supervision of Professor Stuart Madnick concerning such integration[6] . It will describe several different ways of viewing and delineating connectivity and provide a brief discussion of the design of the CIS/TK system. Following this chapter, the reader should have a good understanding of the design and structure of CIS/TK's current state as well as the short-term and long-term goals of the research project.

---

[6] Most of the information for this section comes from the following three papers:

Madnick, Stuart E., and Wang, Y. Richard, "A Framework of Composite Information Systems for Strategic Advantage," Proceedings of the 21st Annual Hawaii International Conference on System Sciences, January 1988.

Madnick, Stuart E. and Wang, Y. Richard, "Connectivity Among Information Systems," Connectivity Among Information Systems, Vol I, September 1988, pp.22-36.

Madnick, Stuart E. and Wang, Y. Richard, "Logical Connectivity: Applications, Requirements, and An Architecture," Connectivity Among Information Systems, Vol I, September 1988, pp. 37-51.

I. In Search of...Connectivity

As we saw in the previous chapter, many companies have tried to achieve an advantage through "connectivity". This is the phenomenon of Composite Information Systems (CIS). The concept of connectivity may be used to describe many different configurations of CIS. I will attempt to delineate the various type of connectivity in two ways: based on the entities involved and based on the actual extent, or "depth" of connectivity (this will be explained below).

A. Connectivity Based on Entities Involved

It is useful to look at the various forms of connectivity, or the different entities that might be "connected", that a company may employ in an effort to achieve a strategic advantage. They include inter-corporate, inter-divisional, inter-product, and inter-model applications of CIS.

1. Inter-Corporate CIS

This type of application involves the linkage of two or more autonomous organizations at some level of their businesses. Examples of the successful implementation of this sort of connectivity abound. They range from the well-known American Hospital Supply success story of the late 70's to the Customer Reservation Systems (CRS's) pioneered by United (APOLLO) and American (SABRE). Each of these "strategic" (at least in retrospect) moves was employed to exploit an advantage in one

of the areas which Porter outlined (refer back to Chapter 1) or, similarly, to limit a disadvantage. For example, the SABRE system for many years gave American Airlines a distinct advantage in booking as they ensured the best positioning of American's flights within the system. While this advantage has since been litigated away to some extent, American still reaps great benefit from the fact that they own one of the two major reservation systems in the world (United is the other).

Thus, by providing a direct technical linkage between the airline and the travel agent, American achieved a significant and relatively sustainable advantage in the (at the time) increasingly-competitive travel industry.

2. Inter-Divisional CIS

These are systems which attempt to tie together two or more groups within a firm. Again, there have been many examples of the successful application of this type of system. Many of these systems have taken the form of automatic order processing by retail branches (examples of this include Toys R Us, Herman's Sporting Goods, and Pepperidge Farms). These examples, particularly the two former ones, show how a company can use such an internal CIS to improve inventory management, reduce stock-outs, improve customer service, and better monitor the sales of their various product lines.

In fact, this can be a first step toward the development of the Inter-Corporate CIS as discussed above. As was the case at Herman's, the success of the internal system caused them to look further into the advantage that such CIS might offer and finally decided to take what seemed to be the next logical step: automatic ordering from selected vendors triggered right from the point of sale!

3. Inter-Product CIS

This type of application involves the combining of systems across product groups. As discussed in Chapter 1, Merrill Lynch's CMA is a good example of this. The CMA could hardly have been launched with such success were it not for the systems support of the the three main products that were combined into one.

Another example, one that will be developed in a great deal more depth in Chapters 4 and 5, is that of CitiCorp's North American Investment Bank. They are currently facing market forces which are forcing them to take a more client-based approach than a product-based approach to their marketing effort. This will involve the connection, at some level, of the many different systems that had supported the various product groups in the past. The problems that they are having in this area will be developed further below.

4. Inter-Model CIS

This type of system attempts to combine various models in an effort to produce a bigger (and assumedly better) model. Examples of this will again arise in the context of CitiCorp where such Inter-Model systems are used to enhance their ability to evaluate potential loans. This type of application, however, is an example of the things being done at another group at the Bank, the North American Finance Group (NAFG).

So, there are clearly many different ways in which a company may find it beneficial to build such composite systems. While they each present the organization with a somewhat different challenge, there are certainly some similarities among them that should be understood. One of these is the dichotomy between physical connectivity and logical connectivity, which is the second of level of categorization, referred to above as the "extent of connectivity".

B. Logical vs. Physical Connectivity

Madnick and Wang have distinguished between these two types of connectivity. The Physical level, or the "first-order issues", which seem to have been the subject of the bulk of CIS-related research to this point, refer to those issues involved with the actual physical connection between the sources. The Logical level, or "second-order issues" are described by Madnick and Wang as "those problems you are faced with once you solve the problems you thought you had (referring to the first-order issues)". These refer to the problems of reconciling the

differences in semantics between the sources as well as inferring concepts that are not explicitly represented in any of the sources.

1. Physical Connectivity

These are the problems that immediately present themselves upon connecting systems with differences such as: different platforms, operating systems, database access protocols, file formats, etc. Within the context of on-line data sources, a good example is the different access of these databases. For example, if while navigating through the database the user decided he/she wanted to go back to the previous section, the command to do this would likely vary greatly between systems. In one, it might be [ESC], while in others it might simply be a "p" (for previous). This is just one example of the many differences that a user must understand in order to perform the integrated usage of the various sources.

2. Logical Connectivity

As stated above, once the physical issues are solved, the user must then face the truly difficult problems which exist on the logical level. The following represents a brief but enlightening example of a <u>few</u> of the many such problems a CIS designer can expect to face:

- **Data Location**: Where are the various data attributes in each database? In a menu-driven system, this entails knowing the

various menu hierarchies, while for relational database systems, it means knowing the table formats of the database.

• **Attribute Naming**: Given that you know what you are looking for and where you can find it, what do the various systems call these attributes? As the example in Chapter 3 will point out, this is a very real problem as not only do sources often differ greatly on what they call common attributes, but the naming schemes are also not always intuitively obvious.

• **Data Formatting and Scaling**: The ways in which the data may be represented will likely differ among (and perhaps within) databases. For example, it is likely that a database will report the revenue of a company will be reported in $ millions. However, other attributes will likely have different scales, owing to their usual orders of magnitude (such as stock price). Further, occasionally databases will present different scaling factors within the same attribute, depending on the particular order of magnitude of that specific value (see Chapter 7 for an example of this at CitiCorp).

• **Inter-Database Instance Identification**: This will likely be a major issue for any CIS: How does one ensure that, for example, company-level data for the same company is retrieved from databases that use different formats (and values) for their company identifiers? While General Motors may be known as "General Motors, Inc." in one database, another might represent it as "General Motors Incorporated, USA." While a person has little

difficulty resolving that the two are the same, the normal computer has no analog to this reasoning capability.

• **Levels of Granularity**: This can be at several levels. For example, at the company level, one database may provide information for General Electric disaggregating all of its operating groups, such as NBC and Kidder Peabody, while others may simply subsume all financials under "GE". Further, at the attribute level, one company may provide detailed financial data through on-line databases, while others provide annual-report-like highly aggregated information. Clearly, comparison between these two companies would be extremely difficult given these different levels of granularity.

• **Concept Inferencing**: Often, the specific attribute that the user is seeking is not explicitly in any of the data sources. However, by using several of them in concert, that attribute might be inferred. The goal is for the CIS to be able to acquire enough information to be able to perform certain levels of inferencing on its own.

II. CIS/TK Design

A. System Overview

Under the tutelage of Professor Madnick, a system has been designed (and a working prototype built) to address these issues. The system is implemented in the UNIX environment to take advantage of its multiprocessing and communications capabilities in order to provide the user with simultaneous access to multiple remote data sources.

As put forth by Madnick and Wang, the goals of the CIS/TK project have been: (1) physical connection to remote databases; (2) DB navigation, attribute matching, etc. ; and (3) advanced logical connectivity. In order to provide these capabilities, the research team has utilized Artificial Intelligence technology (through the use of an object-oriented knowledge representation language) as well powerful DBMS technology.

B. System Design

Please refer to Fig. 2-1 for a graphic representation of the system. There are essentially three levels of processing performed in the system: Application Query level, Global Query level, and Local Query level. To conclude this system overview, I will describe each of these levels from lowest to highest. Of course, the evolutionary nature of the research project precludes any completely precise snapshot of the system. However, this basic outline should represent the general design as it stands today.

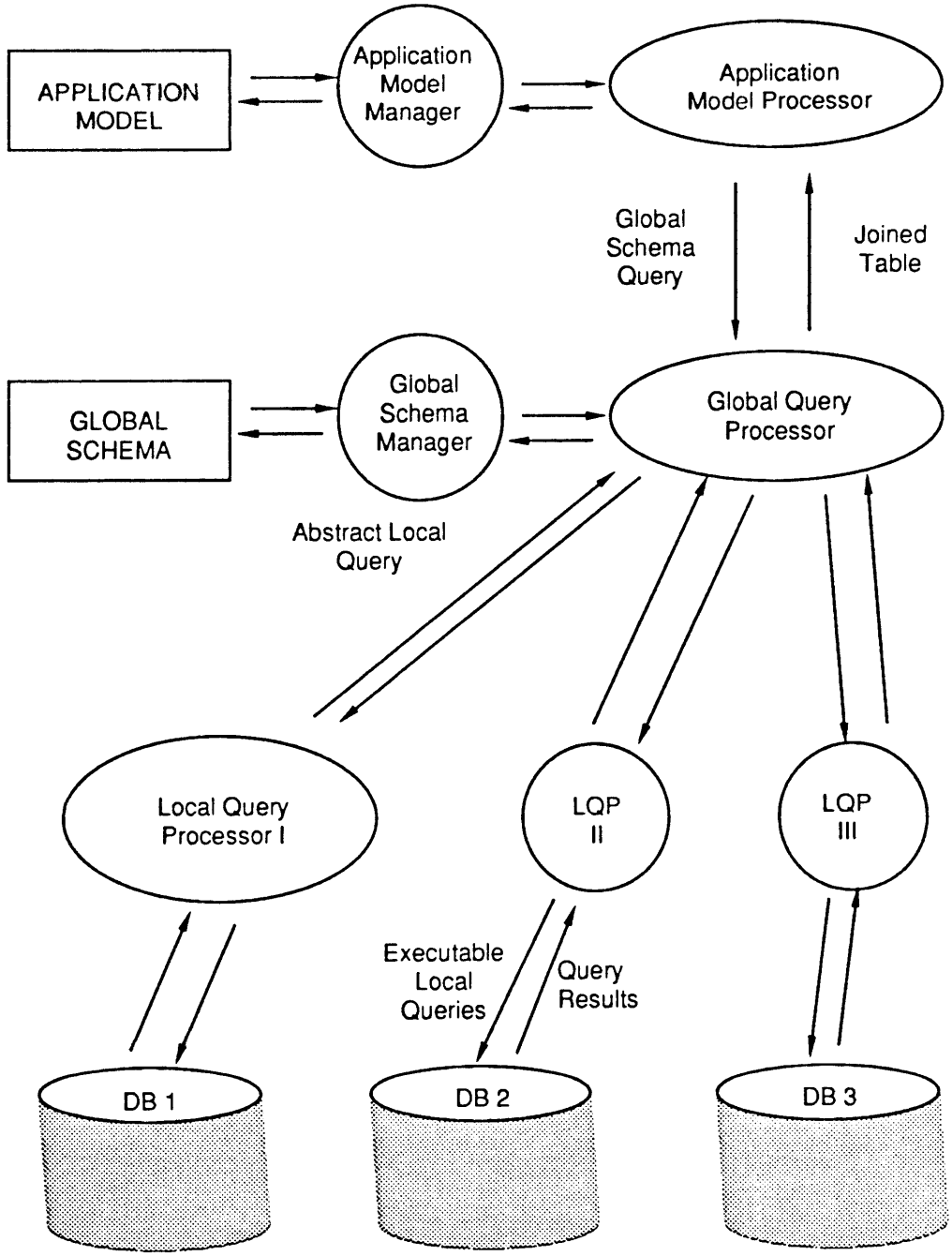


Fig. 2-1: The CIS/TK System Architecture

1. Local level

The term "local" is used with respect to the data sources themselves. The Local Query Processor (LQP) "knows" no more than the schema that exists at the local level of its particular database. A useful analogy might be a company with four or five database people, each of whom is an "expert" in accessing data from a single on-line source. However, there is a coordinator who, while having little understanding of how to get all of the data, knows where everything is. So, when the company needs, for example, the Sales numbers for a company, he hands off this request to the expert (or LQP) in the Compustat (or similar) database (in a language that expert can understand) and he/she performs the "local processing" and returns the data requested to the coordinator.

2. Global level

The General Query Processor (GQP) is analogous to our middleman above. He/she receives a general request for information from one of the "bosses" (each of whom concentrates in a certain functional area) and performs a certain level of "mapping" or translating between the language of the "bosses" into the language of the correct database, given the information desired. Only this GQP/middleman knows about all of the data that is available in this fictitious company. This "knowledge

base" of all of the data available in all of the databases is called the "global schema".

Once the data is returned by the LQP, the GQP again translates it into a language that the requesting boss understands. This processing is likely to include the resolution of many of the logical-level problems that were discussed in the last section.

3. Application level

Finally, at the top level, we have the Applications Query Processor (AQP) which translates the request made by the "boss" into language that the middleman (GQP) can understand. Each "boss" has only a limited knowledge of all of the data available (known as the "Application Schema"), which corresponds to the data relevant to his/her domain, as well as a certain built-in understanding of some of the ways in which he/she would like to interact with that data (such as common calculations or manipulations of the data). This is known as the "Application Model" and is likely to include more data manipulation and logical connectivity tools such as those that perform some level of "concept inferencing".

The CIS/TK team has created a working prototype using this hierarchical design in order to provide the user with the capability of effectively integrating heterogeneous systems and their data. The rest of this thesis will be devoted to documenting the actual needs of users

of multiple sources of data. It is intended for these case studies to provide a direction for the future research and development of the CIS/TK project.

SECTION II: THREE EXAMPLES OF COMPOSITE INFORMATION SYSTEMS

Up to this point, I have been describing the phenomenon of Composite Information Systems, their desirability, and their costs to some extent. In order to better elucidate the concept as well as to confirm the underpinnings of the CIS/TK research project and provide some guidance for its future course, I will now present three examples of situations in which a CIS has been, or will be, used. The first of these examples is from the academic domain and involves an analysis of the October 19, 1987 stock market crash. This analysis makes use of several popular databases that are commercially available and used by many academic and commercial institutions involved in financial analysis. The second example describes the needs of a group within CitiCorp, the North American Investment Bank, in integrating many current standalone systems. Finally, the last example involves another group at CitiCorp, the North American Financial Group, and their desires to integrate data which is to be used in their analysis of marketing opportunities (i.e. new deals).

CHAPTER 3: THE USE OF HETEROGENEOUS DATA SOURCES IN FINANCIAL ANALYSIS WITH A HUMAN DATA INTERFACE

This is a case study which I have performed to begin to investigate the potential difficulties arising in the areas of physical and logical connectivity as related to the use of multiple sources of data in financial analysis applications. It will begin with a description of the topic I have used as an example: a study of a very familiar event. I will then specifically delineate the information that is required to carry out this analysis. I will do this on two levels. First, I will describe the actual financial data which comprised the subject of the analysis (one level of information). Second, on a lower level, I will discuss the "information" or "knowledge" that would be needed to access, integrate, and process this data. Recall the different stages of the problem-solving process as put forth in Chapter 1. This lower level is that information necessary to perform the second phase of the process. The issues comprising this lower level will be further categorized into the physical and logical as defined above. In terms of CIS, it should be reasonably clear that this represents a CIS with a human data interface, as described in Chapter 1. Thus, I will describe the "intelligence" that had to reside at the interface in order to carry this analysis out. This would be similar in that respect to the description of the international bank's data integration operation in Frank, Madnick and Wang [1987].

I. The Problem

The subject of the study is everybody's favorite recent financial disaster: October 19, 1987 (also known as "Black Tuesday"). It was on this day that the Dow Jones average lost a record 22.6% of its total value! In fact, this loss, in percentage terms was even greater than the "other" crash: 1929. The reasons for this most recent collapse have been argued endlessly, and range from automatic computer-based trading to a financial technique known as "portfolio insurance"[7] . Clearly, its effects may very well be with us still in the form of changed market perceptions, and certainly a lower total market capitalization.

The hypothesis that I was seeking to evaluate was the question of whether the disaster had a systematically different effect on different firms. Further, if there was such a differential effect, what were the "sorting factors" along which the effects of the disaster differentiated. To maintain simplicity (perhaps at the cost of significance), I chose as the dependent variable the annualized stock return over two different time windows surrounding the event. As the independent variable, I used several firm-specific, or market-specific, variables that might plausibly affect the way in which a firm was impacted by the crash. Specifically, I looked for main effects for the following independent variables:

- **Size of the Firm:** Of course "size" can be measured many different ways. I chose to conduct two separate studies, looking at the effect of both asset size and income on the lost value.

---

[7] What Caused the Meltdown?" The Economist, December 19, 1987. pp. 65-6.

• **Industry**: Using Standard Industry Classification (SIC) Codes or similar codes, I looked at whether there was a systematically different effect on the firms in some industries from that on others. To further simplify the study, I chose only two industries: computers and automobiles. These industries contained enough variance in the other factors to provide me with fairly range of values.

• **"Market Optimism"**: Using a simple measure such as the price-earnings ratio, I attempted to discern whether the financial disaster redistributed the relative weight that the market placed on current earnings on one hand and the future growth opportunities on the other. Of course, this would take place most likely through the reevaluation of the discount rate which would place a higher value on more current returns.

II. The Intelligent Interface

The interface, in this case the author, must know how to combine this data in order to get the desired outcome. In order to do so, it must "understand" all of the characteristics of this data that were outlined in Chapter 1 and apply that understanding to the specific situation within the context of combining the data with these characteristics. I have defined the "top-level" to mean that information which served as the

input to the analysis itself. It is this lower-level information which is the main subject of this analysis.

A. Top- Level Information

Clearly, on one level, I need to know "what I need to know". This decision would take place during the data gathering stage of the process. To perform this particular analysis, I needed the following data (all at the firm level):

- Time series of daily stock prices
- Information on other adjustments relevant to the market's valuation of the firm. e.g. dividends, stock splits, etc.
- Financial reporting data, including earnings and asset level
- Industry information to be able to differentiate between industries

Two databases were necessary to provide this data in its entirety: CRSP, which provides security-level stock market data on a daily and monthly basis (this includes dividends, stock splits, etc.), and Standard and Poor's Compustat which provides firm-level data based on financial statement information provided on a quarterly and annual basis (more timely updating is actually available, but is unnecessary for much of the analysis performed at the School).

On the next level, I must know (a) how to get that data (e.g. in which databases), and (b) how to integrate and process that data to arrive at the answer I want. In essence, I have already defined how I will process it at the outset. In other words, if there is a statistically significant difference in the value change between, say industries, we might conclude that we cannot reject the hypothesis that there was, in fact, a different impact. The rest of this section will focus mainly on the first question of how I get and integrate the data.

B. Connectivity Issues

As outlined in chapter 2, there are many hurdles that any effort at data integration must deal with at the outset. This problem will serve as a good beginning to understanding many of the practical manifestations of these difficulties and will allow the reader to begin to appreciate the breadth of knowledge that an effective interface must possess.

1. Physical Connectivity

Since this project involved a "batch mode" connection rather than on-line connection, the issues as related to physical connectivity are somewhat different, and perhaps less pressing. For example, the question of LAN's and protocol compatibility, etc. are not really relevant here. Also, both of the databases are in essentially similar formats, and use the same "query language": FORTRAN (although it seems strange to think of FORTRAN as a query language).

Had the databases not been so similar, we would likely have to perform some level of translation on one or both of the datasets. For example, had Compustat resided in a CD-ROM-based Macintosh platform, and CRSP resided where it is, on a tape supporting the IBM VM/CMS system, there would have been substantial physical connectivity problems to be dealt with. However, it is important to note that while difficult, most such hurdles can be cleared, and generally in a great deal less time than their logical counterparts.

2. Logical Connectivity

While physical connectivity may not have posed a great problem to this point, the logical level has clearly pointed to areas in which "knowledge" of each database was necessary to effectively make use of the data. This is a brief discussion of some of the major issues that were faced, and specifically of the "knowledge" that was necessary to make intelligent use of the integrated data.

a. Variable Names

While the content of the databases are very different, there are obviously some overlapping data (without this, it would certainly be difficult - and probably unnecessary - to join the two!). Fig. 3-1 displays some of the data that exist on both and their names in each. It should be clear that to effectively use these two databases, or any combination of any other independent databases, it is essential that one

Description	COMPUSTAT Variable	CRSP Variable
COMPANY NAME	coname	iname(i)*
UNIQUE I.D.	cnum	cusip
TICKER SYMBOL	smbl	itick
INDUSTRY CODE	dnum	isiccd
STOCK EXCHANGE CODE	zlist	iexcd

*the index denotes an array of structures

Fig. 3-1: Same Meanings Hidden beneath Different Variable Names

understand the semantic differences among them while also paying particular attention to understanding where these semantic differences belie logical similarities. This is true as well of semantic similarities which might belie logical differences. An example of this latter case is the existence of different definitions of "Volume" which might exist for different financial products (to be discussed in future chapters).

b. Data Representation

Beyond the relatively simple nomenclature issues exist rather profound differences in the way that the data themselves are represented. A good example is the company name field which is common to both databases ("coname" in Compustat and "name" in CRSP). In Compustat, the name is a simple a 28-character field in which the latest name of the company is stored. However, in CRSP, the name is an array of 32-character "name structures". Without delving into the specifics, each time the name changes in any way (e.g. due to a merger or acquisition), CRSP creates another name structure, while Compustat discards the old name and replaces it with the new one. This is an extremely important point when carrying out an analysis over a long period of time where the names (or cusip codes as discussed below) may, and often do, change (this was not a problem in this study). As shown in Figure 3-2, the names of companies may change very often, even if only in very subtle ways. To ensure that a company is "tracked" throughout its history, it may be necessary to follow these name changes. This is made

Cusip #	Company Name	As Of
03505310	Anglo Lautaro Nitrate Corp	620702
03505310	Anglo Lautaro Nitrate Corp	680102
03505310	Anglo Lautaro Nitrate Ltd.	680715
03505310	Anglo Ltd.	720510
03505310	Anglo Energy Ltd.	801217
03505310	Anglo Energy Inc.	860828

Fig. 3-2: CRSP's Historical Name Change Record for a Sample Firm

more difficult in Compustat as it retains only the recent and the original names.

c. The Unique Identifier

Another important example of a data representation problem is the elusive unique company identifier. This is a major problem for anybody attempting to integrate data across different systems existing at the same level (i.e. company, customer). In each case here, a "cusip" number was used (although it is referred to as CUSIP in CRSP, but CNUM in Compustat). The cusip is the standard corporate identifier (CUSIP = Committee on Uniform Security Identification Procedures). However, in the case of Compustat it is 8 characters long and for CRSP it is 6 characters long (the last 2 representing different security types). Therefore, as this was the join field in this study (and would likely be in any study of its type), a transformation had to be performed which would allow the integrated processing of the two sets of identifiers. Simply, this involved multiplying the 6-digit Compustat code by 100 and testing for whether the CRSP cusip code fell between it and a number 100 higher. Also, a decision had to be made as to how to handle companies with multiple securities. In other words, tracking the stock market performance of a firm may entail following more than one security, such as multiple classes of common or preferred stock.

d. Industry Code

Clearly, to perform this analysis as defined above, I needed to be able to determine in which industry each company competed. For this, Compustat provides a four-character, floating-point industry (SIC) code. CRSP also provides such a code (as seen in Figure 3-1). CRSP's is a four-digit integer code as well, however they do not always mesh. A simple quote from the CRSP manual may give the user initial doubts as to its accuracy: "...The third and fourth digits may not be reliable because CRSP has not verified the SIC codes in any of the files."[8] Further, Compustat has chosen to alter the standard SIC codes for several reasons including "for companies that do not fit any specific classification"[9] (compare this with the way that CRSP deals with such ambiguity: allocating up to five different SIC codes for each company). We will see this problem arise again in the case of the CitiCorp's NAIB where they use the cusip code to identify companies, yet not all companies have cusips. I chose to use one of the two codes (Compustat's) for the main classification number to determine the companies to be included in the study, and then joined those companies with the appropriate ones in CRSP using the cusip as the join field, adjusted as discussed above.

As an example of the extent of divergence among industry codes, please refer to Fig. 3-3 which contains each database's "interpretation" of the types of companies for which I was looking. As shown, the Compustat industry code for the mainframe computer-makers is 3682 and that for motor vehicle manufacturers is 3711. Running a list for each on CRSP

---

[8] CRSP User's Manual
[9] Compustat User's Manual

Car Makers	Mainframe Makers
Code = 3711	Code = 3681
Chrysler Corp	Amdahl
Collins Industries	Cray Research
Federal Signal	Electronic Associates
Ford Motor Corp. of Canada	Floating Point Systems
Ford Motor Co	Prime Computer
General Motors Corp	Tandem Computers Inc
Honds Motor LTD	
Navistar International	
Paccar Inc	
Total = 9	Total = 6

Fig. 3-3: Compustat's Industry Listings for Auto and Mainframe Makers

produced the list in Fig. 3-4. Note the different industry codes. In fact, there were no companies in CRSP with the industry code 3682! Besides several examples of the instance identification problem (i.e. different company identifiers), it is clear that two things are at work: (a) there are some groups not even on the CRSP tape that are on Compustat and vice versa (this will be discussed in more depth below; and (b) each service performs very different categorizations of companies. For example, Compaq computer (3681, or mini- micro computer makers, on Compustat) is listed in industry 7379 in CRSP (which is defined in COMPUSTAT as computer services). On the other hand, Commodore computer (also 3681 in Compustat) was listed in CRSP as 3792, an industry code not used in Compustat. This has very obvious implications for those performing analyses using these databases on specific industrial segments. These mappings from one code to another would (and did) require a great deal of rather specific knowledge at the interface level.

Data Formatting

Of course, it would be utterly inefficient for a database to contain excessive decimal places when the increased accuracy they offered really isn't necessary. Thus Compustat, like many databases, formats all figures in millions ("unless otherwise indicated"). However, this is not the case for CRSP. Specifically, this came into play in my analysis of the price/earnings ratios. This ratio is generally calculated by dividing the price/share by the earnings/share (which can be calculated a number of

Car Makers	Mainframe Makers
Code = 3711	Code = 7371
American Mtrs Corp	Advanced Micro Devices
Checker Mtrs Corp	Amdahl Corp
Chrysler Corp	Anderson Jacobson Inc
Executive Inds Inc	Anelex Corp
Ford Mtr Co Cds Ltd	Applied Digital Data Sys
Ford Mtr Co Del	Barrister Information Systems Inc
Fram Corp	Barry Wright Corp
General Motors Corp	Beehive Intl
General Mtrs Corp E	California Computer Prods
General Mtrs Corp H	Centronics Corp
Great Amern Hldg Corp	Clary Corp
Motor Wheel Corp	Cognitronics Corp
Signal Cos Inc	Computer Consoles
Simca Automobiles	Computervision Corp
White Mtr Corp	...
Total = 15	Total = 60

Fig. 3-4: CRSP's Industry Listing for Auto and Mainframe Makers

different ways, depending on the level of "dilution", or inclusion of non-common stock equity -like instruments, that your analysis - or taste - warrants). Simply taking the total earnings (from Compustat) divided by the number of common shares outstanding (from CRSP) without any adjustment would yield phenomenal EPS figures, and resultant low P/E's which might lead the user toward a perhaps fatally over-optimistic view of the security! The problem, of course, is that CRSP reports shares in thousands, while Compustat reports earnings, and everything else, in millions. The interface, if performing such calculations, must understand these differences in data formatting and adjust accordingly. This is but one simple example of a very common problem in database integration.

f. Intra-Database Data Availability Divergence

The above example points out another piece of "knowledge" that was necessary to perform this analysis. In fact, the EPS figures are available in Compustat, but generally only on an annual basis. Furthermore, the portion of the service to which the M.I.T. Sloan School subscribes only provides Net Income on an annual, not quarterly, basis. This has two implications for this analysis. First, since I knew to "check IF EPS field is filled; IF not, THEN calculate it using the above process". Second, I had to know that "IF the analysis was being performed with quarterly data, THEN the income figure itself had to be calculated using other items available." This understanding of both the data limited by a contract with Standard & Poor's (the provider of Compustat data) and

the reporting tendencies of firms, which determines whether, say, EPS is available quarterly, is essential to the effective use of the Compustat database.

g. Inter-Database Scope Divergence

At the beginning of the analysis, it was necessary to ask the question, "What will be the scope of the study?" Will it be S&P 500 firms, Dow Jones Industrials, etc.? Of course, the analyst is limited to what exists in the available databases. It is therefore essential to understand the differences in the scope of the firms included, since the intersection of the databases yields the only potential candidates. The CRSP files that are readily available at Sloan provide data for only NYSE and AMEX-listed companies (on a monthly basis, about 6,400 securities). On the other hand, Compustat provides quarterly data for about 10,000 companies which are traded on the OTC, Regional, or National exchanges.

h. Reporting Periods

Finally, there was a substantial difference, both inter- and intra-database with respect to the timing of the reporting. CRSP data is, of course, recorded daily (though only updated annually). However, the reporting period for Compustat data depends on the specific company's fiscal year and their desire to report such information on a timely basis. It therefore may be essential when using Compustat data to check that attribute which holds the fiscal year end. This way, the analyst may

make any such adjustments to the analysis such as for earnings announcements, etc.

C. Application of CIS/TK and Conclusions

1. CIS/TK

It should be fairly clear that it is specifically this type of problem for which CIS/TK has been developed. This case study has pointed out several areas which imply that effective connectivity, on both the logical and the physical level, is possible only with some degree of database-specific knowledge which clearly must reside in the interface. Given the importance of such connectivity, it is therefore clear that a system which might act as that intelligent interface would have a substantial impact on the efficiency and effectiveness of such integrated analyses.

CHAPTER 4: THE DATA NEEDS OF THE NORTH AMERICAN INVESTMENT BANK (NAIB)

While the last chapter represented what was essentially an academic analysis and therefore may have less relevance to the data integration needs of the industrial sector, these next two case studies, both from the halls of CitiCorp, will demonstrate that businesses and universities alike have the similar need to combine, integrate, and coordinate multiple sources of heterogeneous data.

This first CitiCorp case study, performed at the North American Investment Bank, demonstrates how the legacy of the earlier "eras" of IT further impede the Bank's march toward a "wired society".

I. The North American Investment Bank[10]

The NAIB, led by Michael Callen, is charged with the mission of providing quality investment banking services to CitiCorp's institutional banking clients. While the difference between investment banking and commercial banking, preserved for decades by the Glass-Steagall Act, has been blurred in the past few years, one important distinction between the two is the fact that commercial banking is essentially an "annuity" business. That is, in the past commercial bankers would extend a line of credit to a customer in return for a periodic interest payment. The investment banking business, however, is more "transaction-oriented" where the intermediary makes money (generally in the form of fees) on the size (and composition) of the one-time transaction.

A. Product Offerings

The NAIB, like most of its Investment Banking counterparts, sells an extremely broad product line to its clients. The list below represents the majority of these product groups:

Foreign Exchange

Japanese Yen
French Francs

---

10 This information was gathered from discussion with the following NAIB personnel: Judy Pessin, Dorothy Conroy, Evan Picoult, John Remmert, Bud Berro, Helga Oser, Dan Schutzer, and Ken Wormser.

Deutschemarks
etc.....

Exposure Management

Interest Rate Swaps
Foreign Exchange Swaps
Foreign Exchange Options
Caps/Floors
FRA's
Fixed Income Options
Exchange Futures & Options
Options on Futures
Swaptions
Investment Agreements

Securities Distribution

Bills
Short Coupons
Long Coupons
Agencies
Zero-Coupon Bonds
Foreign Debt

Debt Origination

Mortgage Backed Securities

Finance

Money Market Instruments

Long Term Finance

Municipal Finance

It is crucial to remember that this is simply a snapshot in time. Any description of their product line is good for only a short while given the

volatility and competitiveness in the market and the resulting rapid product development cycles necessary to remain a major player.

Given their tremendous financial assets and expertise, they also engage in trading on their own account in these same financial markets to which they provide access for their clients. This trading function comes under the responsibility of the NAIB as well. Given the inherent instability in the trading, it is desirable to have a large portion of income coming from the more consistent client transaction fees. Currently, the split between such transaction fee income and trading income is 30:70, while they are hoping to improve it to closer to 50:50.

The changing economic landscape faced by CitiCorp, as discussed above, has led to a major change in thinking. Under Walter Wriston, their culture has been decidedly decentralized, and his successor as CEO, John Reed, has continued this to a great extent. However, the increased need to focus on customers rather than products has resulted in a slight shift toward relatively more decentralization. Still, on an absolute scale, one would still consider CitiCorp a Bank with an extremely decentralized culture. On the systems side, the result of this type of structure is an widely-distributed systems environment to which the "stovepipe" model presented in Chapter 1, truly applies. In fact, just about each of the product groups mentioned above runs on its own system. More on this later.

The industry has become more competitive. Their product line has grown and grown in response. There has developed a need for CitiCorp to integrate its dealings with customers across its product line in an

effort to develop more of a "relationship" with each customer. Given these factors, along with the intensity of technology in the value chain of most financial services players, it is not surprising that this has led to the need for the technical integration of some sort to support their move toward business integration. This example stands as a classic example of information systems which were not made to work together (for organizational reasons) yet are now being asked to do so.

B. Organizational Structure

The NAIB operating entities relevant for this analysis are: Risk Management, Credit, and Profitability. These are essentially the main "user groups" of the current information systems at the NAIB that are in need of integration. This integration effort is further complicated by the fact that these various constituencies are looking for very different functionality from this integration, as will be explained shortly. To get a general feel for the relationship of the client (or investor), CitiCorp, the various product groupings and their systems, and these three functional areas, refer to Fig. 4-1. This clearly demonstrates the complexity of these relationships and the difficulty of an integration effort involving over 10,000 accounts, up to 10,000 transactions per day for some products, and 20 different systems.

II. The Users of Integrated Data

This section will describe in more depth the roles and responsibilities of these three functional groups. Further, I will discuss their data and systems needs and begin to develop the presentation of their connectivity problems, which will be laid out in the next chapter.

A. Credit

In 1981, CitiCorp became the first Bank to add to their traditional investment banking control process a Credit function. While this was certainly not a "new" idea in financial services (in fact, it is the credit function which acts as a filtering device through which most financial services transactions are screened to meet bank and regulatory guidelines), it <u>was</u> new in investment banking. The impetus for this innovation, like most made by CitiCorp, was decidedly market-based. At first glance, the investment banking business seems to have little need for a credit function: there is not the typical extension of credit in return for a promise to pay in the future which characterizes most of the other transaction-oriented banking practices (typical of commercial banking). However, during the early 80's and late 70's, there were several banks that experienced failures resulting from clients committing to future transactions and then failing to "deliver" (either completing their side of the buy or sell transaction as the commitment specified). The loss to the investment bank in this situation is the difference between the value of the securities on the date of failure and

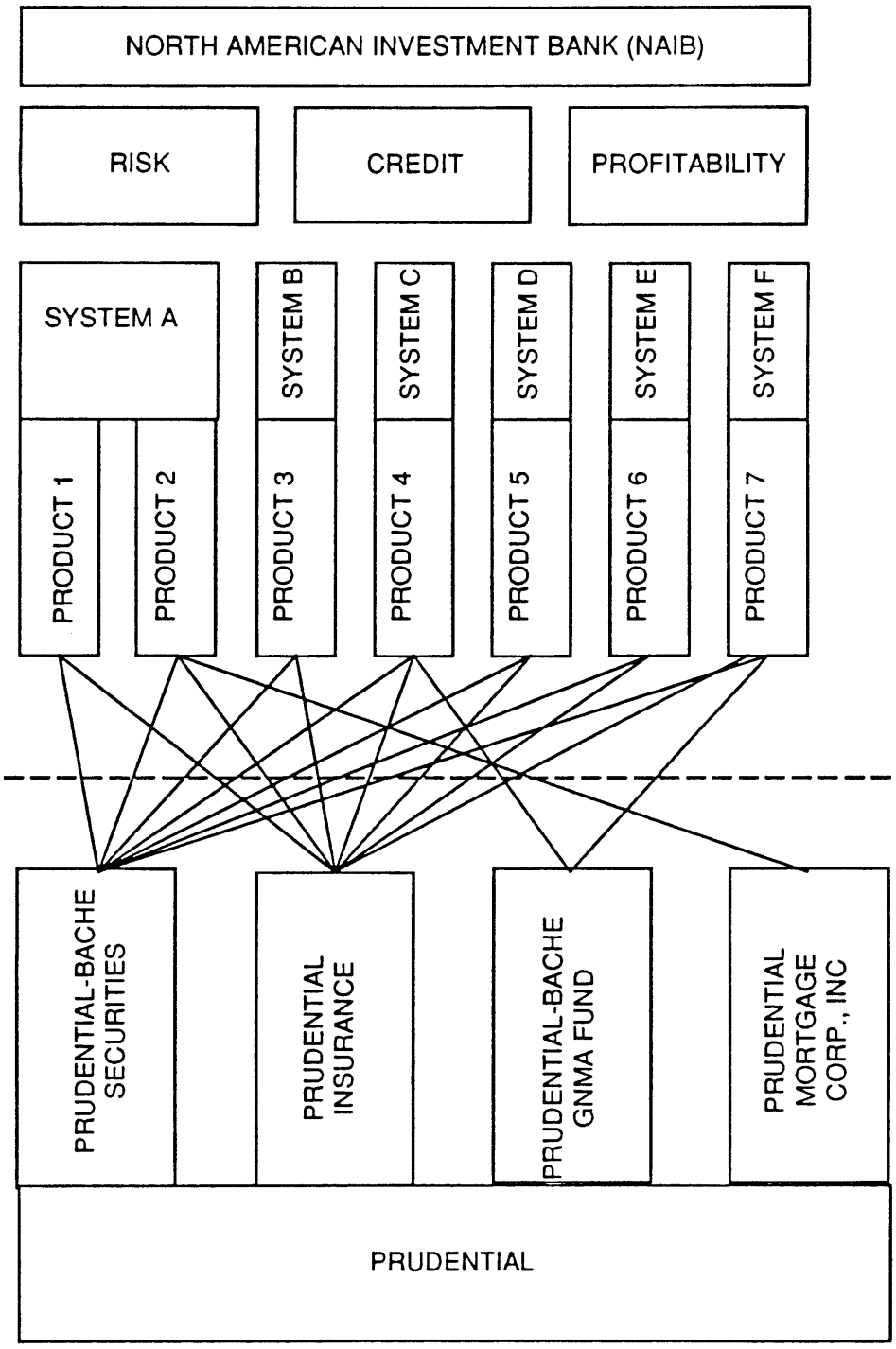


Fig. 4-1: The Complexity of Relationships at the NAIB

the agreed upon value in the commitment (of course, depending on the way the market happened to move in the intervening period, there might actually be a windfall gain to the bank as a result of the failure to deliver).

Essentially, the risk arises in this case from three sources: (1) the size of the temporal window between the initiation and commitment of a future financial transaction and its eventual execution and delivery; (2) the volatility of the specific financial instrument under consideration; and (3) the creditworthiness and financial strength of the counterparty.

In order to manage this risk, CitiCorp's Credit Department monitors daily the "exposure" that the NAIB has within each product vis-a-vis each client. An example of a report that the credit manager would look at is shown in Fig. 4-2 (note that this is for one single product). These and similar reports present calculated data that estimates the loss to CitiCorp were that client to declare a failure on the current day. The data is presented first by product (since it comes off of each product's system), and then by client within each report.. These reports are then used to control the amount of business that may be transacted with a given client based on their total exposure, the riskiness of that exposure, and on their inferred financial strength.

The data that serves as an input to the Credit process consists of the internally-produced transaction data representing the open trades that exist (the forward commitments) as well as hard and soft data concerning the client's financial health. The credit manager then integrates this data, processes it and attempts to ascertain the true

Fig. 4-2: A Sample Credit Management Exposure Report

Trade Date	Settle Date	Security	Amount ($MM)	Price			Expos. ($M)
				Contract	Mkt	Change	
3/23/87	4/8/87	GNMA 9.500	9.80	95.31	94.62	.69	+67.6
3/25/87	4/11/87	GNMA 9.000	-1.00	91.25	92.03	.78	+7.8
3/26/87	5/1/87	GNMA 9.000	.30	91.53	92.03	.50	+1.5
3/30/87	5/2/87	GNMA 9.500	-.42	94.40	94.62	.22	+0.9

expected value of the firm's future transactions with this client, and do its part to maximize this value. The "answer" to his/her analysis would be a decision as to the advisability of future business with a specific client, the setting of a new product- or client-level limit, or perhaps the conclusion that everything is "Okay."

So, Credit is charged with the responsibility of monitoring many different companies which often do business with CitiCorp across many different product groups. Credit is clearly a client-level function. For the credit managers to adequately perform their task, they must process financial information that they have gathered at the client-level (and to a lesser extent market-level macroeconomic data) along with internally-produced data which is organized at the product-level (primarily, with the client as the secondary level of aggregation). Within the context of heterogeneous data sources, then, we see three interfaces in which different types of data are combined. In each case, the interface, whether it be a human, a machine or some combination thereof, is challenged to deal with various data differences in order to perform the necessary analysis:

- **External-External Interface**: How does the credit manager integrate all of the external sources logically? How does the credit manager deal with a Wall Street Journal report that says that a company is in a great deal of trouble at the same time as hearing "on the street" that the company is in the middle of a major turnaround that will produce excellent results (a contradiction in value)? Clearly, the past experience with, and credibility of, each

source is one key determinant of the relative weightings of the various data.

- **External-Internal Interface**: This is obviously where the credit manager really adds the value. How do we combine the information telling us that Client Z has just laid off 1,200 employees at its Kenosha, Wisconsin plant with the fact that this client has $X million of outstanding trades with CitiCorp's foreign exchange business? Further, how does the Credit manager identify the external data as being related to the same company as the internal data (the instance identification problem)?

- **Internal-Internal Interface**: This is the area of the NAIB where CIS/TK seems to be the most applicable at its current stage of development. How do we combine the information on the exposure (as defined above) of Client Z in the foreign exchange area with exposure in caps/floors? How do we do this logically as well as physically? This integration is clearly essential for effective credit management and <u>is</u> done currently at some interface, almost exclusively human. One goal is to offload this responsibility to a more consistent, rapid technical interface.

Thus, the goal is to transform data aggregated at the product level int data aggregated at the client level. It seems in some sense that this is a simple example of a problem that the relational database was designed to solve. That is, rather than preprocessing the information (i.e. representing it in the system at a higher level of aggregation- either client <u>or</u> product), why not just put it into one big database and then

users would be able to use a RDBMS to specify at what level of aggregation they would like to look at the data (presenting customized "views" of the data)? The answer is that even if you can technically do so, but you shouldn't. Refer back to the Management in the 1990's framework (Fig. 1-3). The argument for one big database ignores (at least) the organizational factor. It is CitiCorp's stated strategy to maintain autonomous product groups. Given that, along with the political "turf" issues that tend to arise with respect to the ownership of data and various other factors represented by the nodes in the model, it is essentially a foregone conclusion that the notion of a "company-wide relational database" is impossible in this situation (and probably many others).

Strategically, NAIB's long-term goal is to be able to provide the client with a single line of credit rather than up to 20 credit lines (depending on how many of CitiCorp's product areas in which they do business). This goal, however, will likely be impossible without the full integration of data using a technical interface.

B. Profitability

As its name suggests, this department's concern is making money. Specifically, they are concerned with how, where, and how much the NAIB does so. Their focus is at three levels: (1) Client, (2) Salesperson, and (3) Product. The latter level will be ignored since the data is all aggregated at that level and the determination of profitability at that level is therefore relatively straightforward.

The importance of this data cannot be overestimated. Like any provider of a multi-product line, CitiCorp has a variety of pricing schemes which are geared toward maximizing overall corporate profitability rather than at the product level. The result of this is that some products are priced extremely low in order to attract clients who may be more likely to purchase the more complex and therefore more expensive products (based on their inferences of inter-product elasticities, etc.). As a result, CitiCorp needs the capability to monitor client-level profitability. At the extreme, they may want to terminate a client relationship if the preponderance of their purchase are concentrated on the "loss leader" (or simply low-priced) products, resulting in a low, or negative, contribution to CitiCorp's profits. Similarly, they want to be able to identify cross-selling opportunities between the various product groups that meet similar, or complementary, needs. This complementarity can only be revealed through the analysis of such integrated data.

On the other hand, CitiCorp has a strong interest in ensuring that their salespeople do not sell these lower-profit (often easier-to-sell) items while at the same time forsaking the big-ticket products. This requires salesperson-level profitability analysis. In addition, CitiCorp is currently in the middle of implementing a "team selling" program for their big accounts. This will involve salespeople for government securities, options, foreign exchange, etc. coordinating their efforts and, in so doing, providing a single interface between CitiCorp and the client.

1. Evaluation of Salespeople

At the salesperson level, it is very difficult to effectively determine the relative value of a selling effort across the heterogeneous product lines. For example, if salesperson A sells $100 million in T-bills and salesperson B sells $50 million in foreign exchange options, who has done a better job? One approach would say that the more valuable product, that is the product which attracts more income to CitiCorp, would be assigned a higher value. This, however, may be a rather short-term view (depending on the nature of the product as well as its place in the product line) and may ignore long-term returns (as well as opportunity costs).

Currently, the NAIB is implementing a "sales credit" program which attempts to assign various values to different products. So, for example, selling the $100 million in T-bills might net me 45 selling credits, but the $50 million in options might earn me 75. This takes into account the overall value of the product, the difficulty of the product to sell, the complexity of the product, etc.

As is clear, to implement this evaluation program, at the individual level (it would be even more difficult for the teams), there is a need to aggregate the data at the salesperson level. Currently, this is done manually, with each system calculating and outputting the sales credits that each salesperson earned. So, it is again necessary to take the product level data, which exists in many different locations, and aggregate it at a different level, this time at the salesperson level, to allow for the tracking and documentation of the total sales production of

these multi-product salespeople. This presents the interface with some interesting challenges which will be discussed in the next chapter.

2. Investor Level

It is becoming increasingly important, particularly with the advent of "investor teams", to understand CitiCorp's profitability vis-a-vis each investor for whom they provide investment banking services. While it may seem straightforward to identify the performance of CitiCorp at this level, it is not so at all. Again, this involves another level of aggregation, or another "link" that must must be created. Refer back to the example in Section A where the Credit department is aggregating product-level data at the client level. This is exactly the same problem here. However, even within the client-level, there exist different degrees of aggregation (i.e. departments, subsidiaries, legal entities). To complicate the problem, to some extent the level of client-level aggregation needed to support Credit's business will not be the same as that necessary to support Profitability's business. More on this in Chapter 5.

C. Risk Management

The third, and final, entity which is relevant for this analysis is Risk Management (or simply "Risk"). Their function is to monitor the exposure of CitiCorp to the various market and macroeconomic factors.

They are mainly concerned with such measures as the interest rate risk and foreign exchange risk of CitiCorp's held portfolio of securities.

The data needs of Risk provide an interesting difference from those of the other two groups discussed above. While Credit and Profitability were each concerned with re-aggregating the data in a different way, Risk is concerned simply with the total, or the bottom line. That is, Risk is concerned with the ultimate level of aggregation: CitiCorp.

To perform the risk management function well, it is essential to have the following information (1) an understanding of the sensitivity of various products or securities to changes in, say, interest rates (a "model"); (2) an understanding of the current portfolio of these products (internal data); and (3) an understanding of where the macroeconomic variables are today and where they will be in the future (or at least some estimation to that effect).

It is essential that this data be aggregated due to the complex interactions among the various products. For example, a treasury bill's sensitivity to the general level of U.S. interest rates would tend to be negative (i.e. a rise in rates would tend to drop the value of a held t-bill). However, a Yen-denominated call option on the dollar may react positively due to the increased relative attractiveness of dollar-denominated investments precipitated by the rise in interest rates. This contrived example involving only two products should help explain the complexity of the risk management process and the need for total integration of data on all securities held.

The analyses performed on this data are then mainly used to set limits on the activity of traders in the various markets. This is the only way they can effectively control the Bank's risk. This points out the very real need for accurate real-time market data in order to react swiftly to market changes. To grasp the importance of this information (particularly that related to CitiCorp's portfolio), in 1987 Merrill Lynch lost $250 million due to the activity of a single trader in a single day! While he did nothing illegal (or unethical), it is clear in retrospect that he took a far too risky position for the given product. It is precisely this type of situation that Risk Management is charged with the task of avoiding. However, the opposite situation is also to be avoided. That is, traders consistently taking too little risk would lead to a less-than-optimal use of CitiCorp's resources. Therefore, "Risk Management" should never be construed as being synonymous with "risk minimization".

This brings up another important point: the necessity for historical data. The role of historical risk data would be as a feed into the setting of the current limits. Its role would be that of allowing the Risk manager to ask questions such as "what is the relationship between the risk taken by a trader and his/her realized return?" By understanding this, the CitiCorp risk manager will be able to more effectively monitor and control the level of acceptable risk taken by traders.

As will be discussed in some more depth below, this is one area in which there has been a certain level of attempted technical-interface integration. A system called "Utopia" was designed to provide the risk managers with data from all of the product groupings and the current

inventory/maturities etc. of each. It presents an excellent example of a combined human-technical interface and seems to have worked very well for the Risk managers. In Chapter 5, I will describe the system in more depth as well as point out the very real differences between it and a system such as CIS/TK in the level of true logical connectivity they provide.

CHAPTER 5: SYSTEMS AND DATA INTERFACES AT THE NAIB

The last chapter outlined the basic functions performed by Risk, Credit, and Profitability as well as the data that is necessary for them to carry out these functions. This chapter will now go into more depth on the specific problems that will be (and have been) associated with providing the NAIB with the level of integration for which each of its groups is looking. This will provide more concrete examples of the different problems associated with physical and logical connectivity as well as providing more support for the notion that the logical level is the area most in need of immediate attention as there are few hard "answers" to the problems it presents.

At the most basic level, each of the three groups are data processing functions. Of course, they are far more than the technology-intensive number-crunching computer centers that we generally associate with the term "data processing", however they each take as an input multiple streams of data, process it and analyze it, and output more streams of very useful data and information. As Fig. 5-1 shows, the nature of the inputs, processing, and outputs differs greatly across the groups but the general structure is common to all three and conforms to the paradigm that I put forth in Fig. 1-4.

It should be clear that computer systems can and have been used in each of the components of this diagram. It can be used to both collect data for their processing (such as from on-line databases like Reuters) and disseminate the final data (such as the notification of a change in trading limits on an E-mail type system). The integration of heterogeneous databases, however, mainly occurs in the middle portion, the processing (and perhaps the analysis) of the data that has been collected. Of course, depending on what one considers "gathering" vs. "processing" data, he/she might consider part of the integration problem to be contained in the first stage (such as the issues related to physical connectivity).

I. Systems at the NAIB

Like most decentralized organizations, CitiCorp planned its systems design effort around its lowest relevant autonomous units of operation.

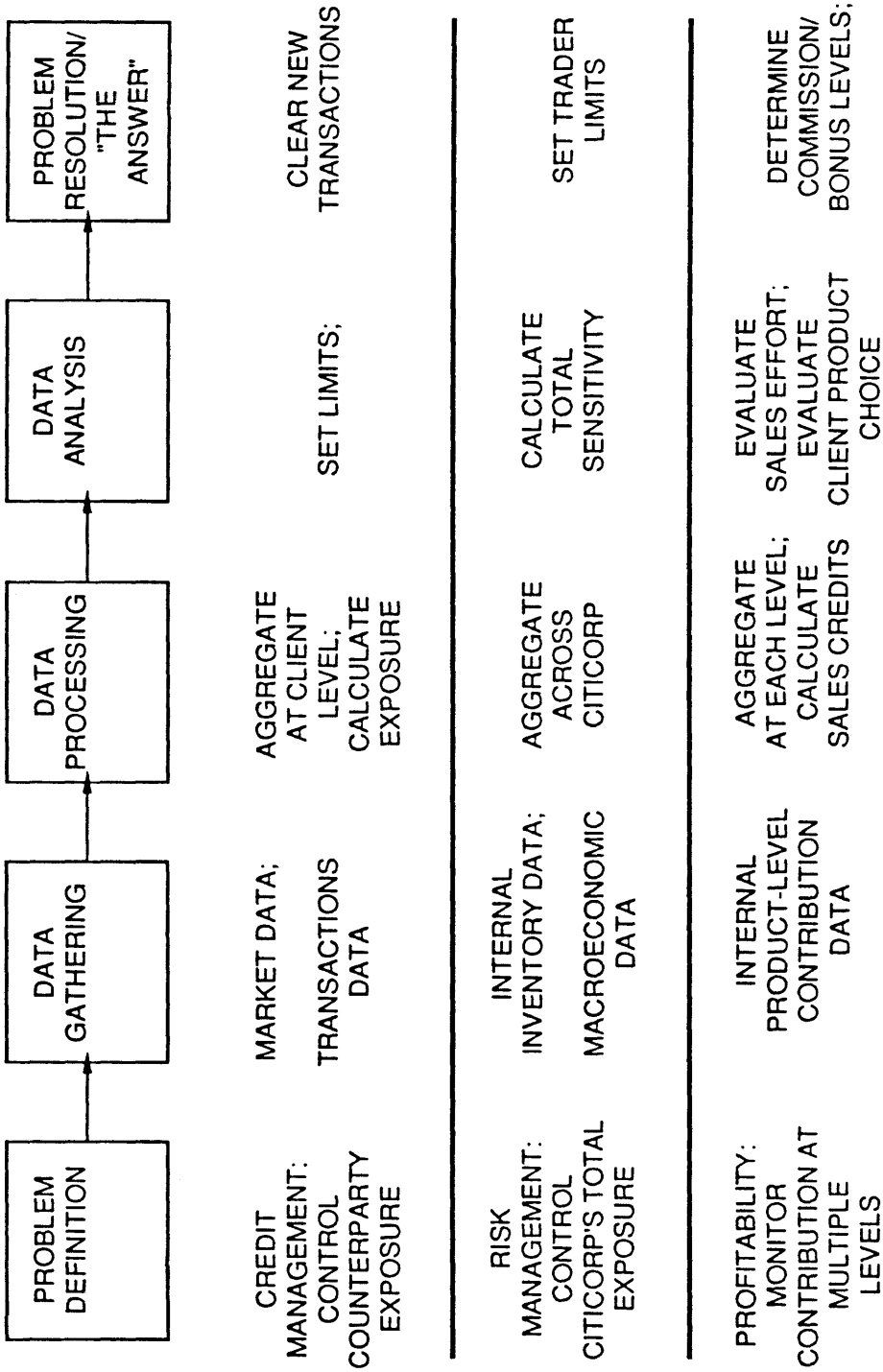


Fig. 5-1: Nature of the NAIB Problem-Solving Processes

As outlined, CitiCorp is extremely decentralized company in which each unit is often judged just as any external entity would be: on its bottom-line (after allocation of costs, etc.). The result of this has been the evolution of many stovepipe systems where vertical applications have been developed, whether in-house or externally, that are dedicated to specific products or product groups.

In fact, NAIB alone has over 20 different "systems" , which corresponds roughly to the number of product groupings offered. See Fig. 5-2 for an outline of many of the specific systems, their platforms, and their processing responsibilities. Note particularly how many different platforms on which these heterogeneous systems are run. This reflects the very real differences in tracking and processing needs that exist among the various product groups as well as the organizational goal of a high degree of autonomy. The rest of this chapter will be devoted to outlining the issues involved in order to meet the data needs of the NAIB groups as outlined in Chapter 4 within this environmental context. It is interesting to note that this problem is likely to have occurred with or without tremendous foresight. That is to say that it is in no way due to an error of judgement along the way that CitiCorp is facing this difficult situation. It is simply the systems ramifications of the "other two legs of the table": strategy and organization.

In a stable, fast-growing, fragmented market, the distributed system configuration may in fact be very desirable. Particularly at a company like CitiCorp, autonomous systems are desirable for several reasons:

Processing System	Accounts	Hardware
CTS	10,000	IBM 3090/MVS
CitiTracs	10,000	IBM 3090/MVS
CRA	1,000	DEC
IPPS	225	IBM 3090/VM
Devon		IBM 3090/MVS
Asset Sales		Prime
Masterpiece	175	IBM 3090/VM
RealTime		Wang
Microbook		Wang
Brennan	500	DEC
Q Swaps	683	Prime
Futrak		PC
Transaxis	1,500	DEC
Currency Trader		IBM

Fig. 5-2: The NAIB Processing Systems

- **Corporate Culture**: John Reed's view of CitiCorp is still very much as an affiliation of small businesses, in an effort to harness the entrepreneurial energy of small groups. It would be rather difficult to preserve this feeling in an environment where, similar to Ford Motor Corp in its early years, "You can do anything with your systems as long as its the same as the rest of the bank."

- **Break-Up**: In such an affiliation of small businesses, it will inevitably become necessary over the course of time to divest oneself of certain businesses and acquire others. Under the "one big system" idea, this may become a difficult thing to do. Several companies with major centralized systems have found this to be a stumbling block in divestiture efforts: the value of the divestable unit may be severely impaired by the lack of portability of its processing system

- **Needs**: It is highly unlikely that a high-volume transaction product like government securities trading will have even remotely similar system processing needs to those of a more complex, "deal-oriented", product such as caps/collars. In essence, a system which tries to be all things for all products is doomed to failure.

Recognizing these facts, CitiCorp has built a multitude of systems that were simply not meant to work together. Now however, their strategy and the evolution of the industry (as outlined above) has dictated the need for just such a level of integration.

II. Meeting the Data Integration Needs of the NAIB

While they are each similar in some respect, and interdependent in many respects, the three groups will be treated separately in this analysis (with appropriate cross-references). The actual integration effort will undoubtedly reveal a great deal of overlap and will surely be done in concert, however their differing needs will be elucidated more easily under this separate treatment.

A. Credit's Data Integration

As outlined in Chapter 4, Credit is concerned with taking product level data and re-aggregating this to client level data in order to match it across product groups as well against the client-level exposure data for processing and analysis. The main issues that they face in solving this quagmire involve semantic differences, instance identification and client entity identification. Each of these will be discussed in turn following a brief description of the current integration process.

1. Current level of Integration

Given the obvious value of integrated data to the Credit department, it is no surprise that they have made efforts to present the credit managers with some level of integrated data. However, as of today, it is

done manually. So, each day, a person (the human data interface) compiles a report for each client that combines the exposure data output by each product system. CitiCorp recognizes that it is clearly desirable for this to be done using a technical data interface to at least some extent. The repetitive nature of the task, as well as the need for accuracy seems to cry out for a technical solution of some sort.

2. Expected Problems with Credit's Logical Connectivity

a. Semantic/Formatting Differences in Credit Data

Refer back to Chapter 1 to the discussion of the characteristics of data. These concerns refer to differences and commonality in the meaning and differences in the formats among the various data. The resolution of such problems is crucial for any inter-database interface. They represent some of the "knowledge" that the human interface undoubtedly possesses and processes in order to compile the aggregate exposure reports. A good way to grasp some of the issues is to look at the exposure report shown in the last chapter (Fig. 4-3) and realize what it is like to aggregate this across product groups. Doing so, we can unearth some clear examples of the problems that are faced by this data interface:

> **Positive/Negative Exposure:** It is somewhat difficult to understand the sign convention used to describe "exposure" of CitiCorp. Our understanding is that this has changed over the years with a positive exposure currently meaning that there is a

potential loss to CitiCorp. This seems to be one of the those pieces of data that "one just knows because he/she knows." However, this may not always be the case. It was brought to our attention that in at least one case (immediately following the stock market crash in 1987) a credit manager had to inquire of the systems people as to whether a positive exposure figure on a specific product was good or bad! It is unclear whether there is a common standard across systems for this convention.

Settling Period: Credit managers have resigned themselves to the fact that they can hardly monitor every outstanding forward commitment made by CitiCorp. Therefore, they have created a delineation of trades into "Cash" trades and "Forward" trades. The variable which distinguishes these two is the time until the expected delivery of the trade. This measure is used as a surrogate for the credit sensitivity of the trade. However, the sensitivity of, say, a 60-day forward commitment differs greatly between a T-bill and a Mortgage-Backed Security. Therefore, the cutoff for a trade being categorized as "cash" (and therefore being ignored by the credit department) differs across product lines. So, for effective integration of the data, the "interface" must understand these differences.

Formatting/Scaling: This is one of the most common problems that is faced by a company trying to fully integrate data that already exists and CitiCorp is no exception. There are actually two levels at which this problem exists: (1) Intra-Product: Even within the single product reports, there are different scaling

factors used. Of course, this is due to the different magnitudes of the data contained therein. For example, the par value of the trade is carried in $ millions while the exposure is carried in $ thousands. Obviously, this is due to the fact that the exposure tends to be a small % of the total par value. However, it is crucial for the interface to understand these differences when integrating the data. (2) Inter-Product: Again, owing to the different magnitudes associated with the normal transactions of each product, there are different scaling factors used and any attempt to arrive at a "bottom-line" number from this list of numbers must first take into account the scaling factors used here.

b. Instance Identification

As was outlined in Chapter 2, the instance identification problem is a common one in the integration of heterogeneous databases. Particularly when dealing with the large customers who do business along the entire product line of CitiCorp, any one client might have up to 20 account numbers (actually more, due to the various numbers of entities that may be affiliated with any one client, as will be discussed below)! It is obvious what a problem this would cause when trying to aggregate at the investor level.

As it is currently integrated manually, the integrator has to "know" all of the product-level account numbers of all of the clients under study. The technical interface would need to be able to map the name to the various account numbers. As an example of the complexity of this

problem, refer to Fig. 5-3 for an example of a fictitious large company. The completely unrelated numbers as well as the different formats pose a significant mapping problem for the interface.

There has been some movement for the standardization around the cusip identifier of the firm. One problem with this, however, would be that the cusip is a security-level measure. Therefore, any firm that has not been assigned such a code must be given a sort of pseudo-cusip code. This then has the potential of causing very real problems down the line (for example, if one of the pseudo-codes is ever used by an actual new security). There is a great deal of attention being placed on the resolution of this problem at CitiCorp. It is further complicated by the multiple entity levels as described below.

c. The Entity Problem

As I briefly mentioned above, a large company which does business on all of the 20 systems of CitiCorp may actually have a great deal more account numbers than simply twenty. This is due to the fact that many different "entities" within a company may be doing business with CitiCorp. See Fig. 5-4 as an example of how difficult this can become. This list shows all of the entities that were on the main CTS system for one client: Prudential.

The question that is important for those in Credit to answer is at what level is their analysis most important? That is, at what level of the company (i.e. parent, holding company, fund, etc.) should they be most

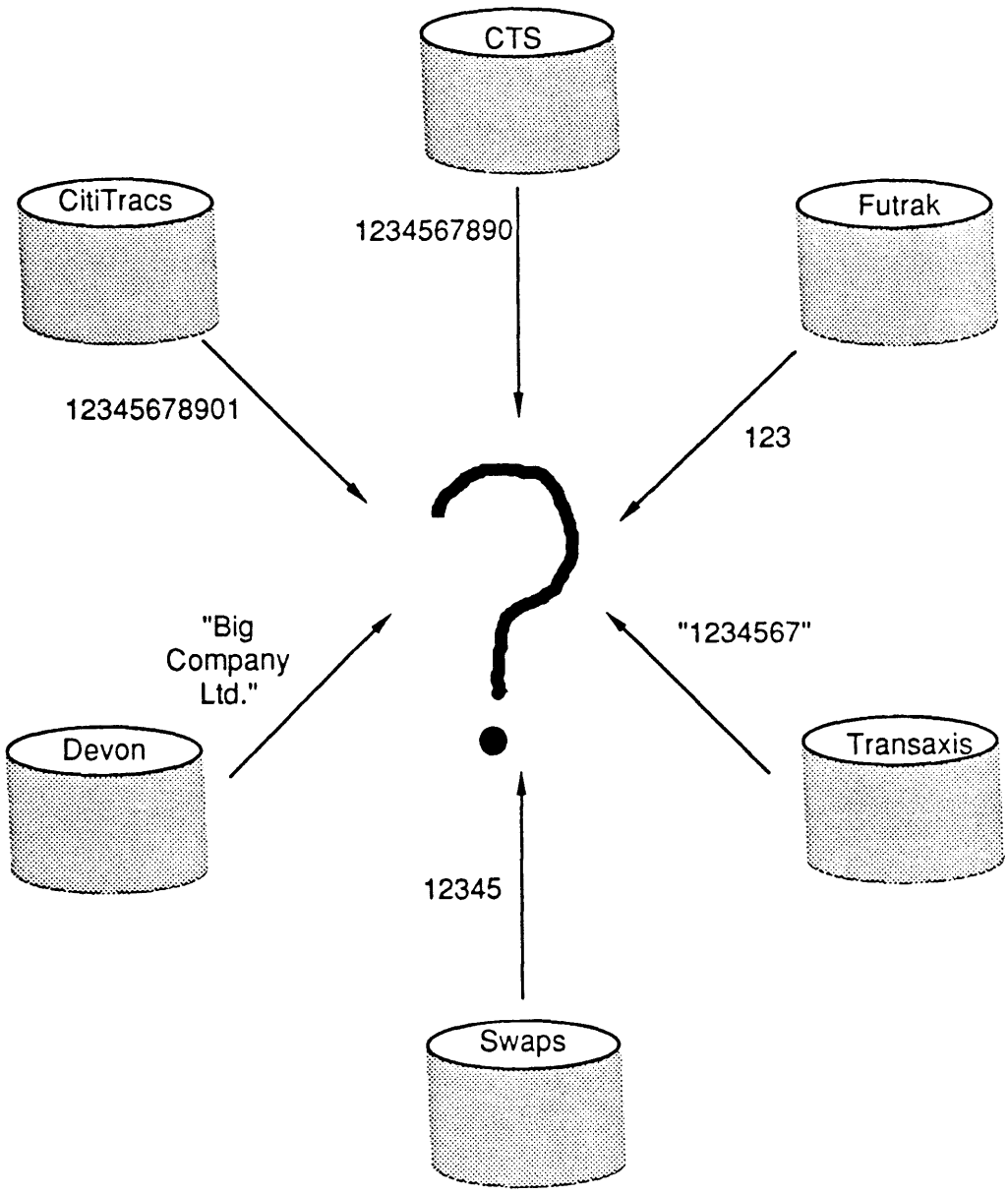


Fig. 5-3: The NAIB Instance Identification Problem

Level I: Prudential

Level II: Insurance
Broker/Dealer
Money Funds
Financial Services - Other

Level III:

Insurance:

Commonwealth of PA Global BD
Extended Reinsurance Group
May Carter Assoc.
Metro Knox Solid Waste Authority

Plus additional 35 insurance entities (accounts)

Broker/Dealer:

Prudential-Bache Securities, Inc.
Prudential-Bache/Puerto Rico

Funds:

Prudential-Bache GNMA Fund, Inc.
Prudential Liquidity Port Money Mkt. Seriea
Prudential Strategic Income Fund
Prudential-Bache Govt Plus II
Prudential-Bache Global Fund
Plus 29 additional funds (accounts)

Other:

Prudential Mortgage Co., Inc
Prudential Funding Corp.

Fig. 5-4: The Various Entity Levels of a Sample Customer

concerned with gathering financial data and evaluating financial integrity? They tend to lean toward the analysis of the "legal entity" level (which, in Fig. 5-4, is represented by Level III). Technically, these are separate operations which could go bankrupt or have other financial problems from which the rest of the firm is insulated. Therefore, they assign each account two numbers, a "credit account number" (which is the legal entity's identifier) and an "account number", which may coincide with the credit account number, but is more likely to be one of several accounts which are tied to a single credit account number. For example, while XYZ Corp. may be a legal entity, it could have account numbers for its pension fund transactions, its international hedging transactions, its cash management, etc. These are all then mapped to a single "credit account number" which aggregates the product-level data for the legal entity.

While this level of integration is performed, the problem is doing so across systems. I get the clear impression that the assignment of various entity statuses is not consistent across the systems. For example, while XYZ Corp. may have three account numbers (i.e. legal entities) in CTS, CitiTracs might have a very different list, with some entities that are not included on CTS' list as well as combinations of certain entities on CTS into one large entity. This problem is represented in Fig. 5-5. This different level of granularity presented by each system greatly increases the difficulty of integration at the client level.

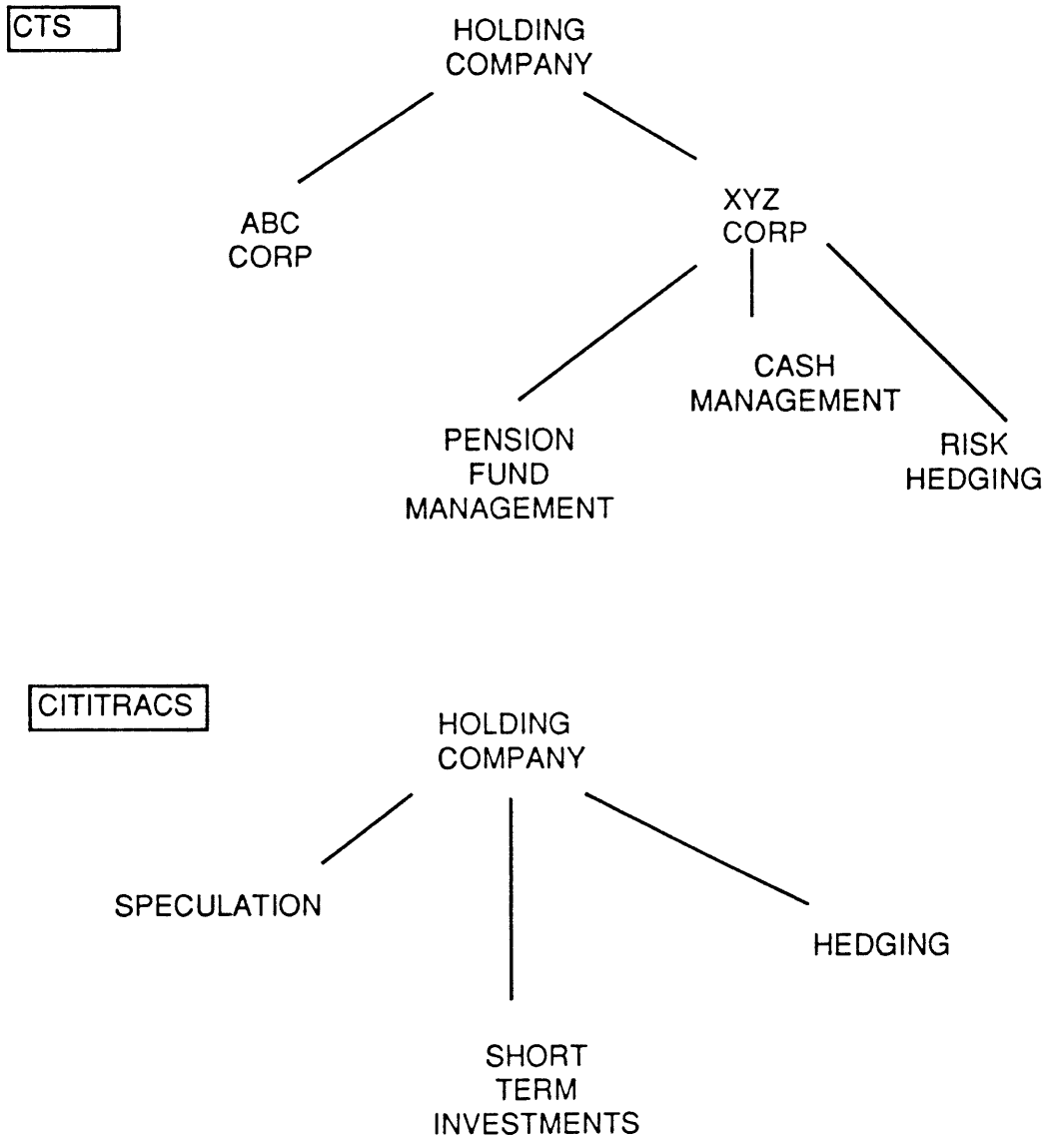


Fig. 5-5: The Entity-Level Problem at the NAIB

B. Data Integration for Profitability

While the general nature of many of the issues are the same as those being faced by Credit, the different uses that the data is being put to, and the different levels of aggregation which are relevant pose additional challenges to the data interfaces employed by Profitability. The problems for profitability, while having different manifestations, come under very similar headings, the two most important being (1) Entity questions; and (2) Instance identification.

1 Entity Questions

While Profitability certainly faces a similar problem to that of Credit in terms of the various levels of entities on which they could aggregate, the problem is made more complex by the fact that the levels that they are interested in are likely to be different from those that Credit is interested in. Therefore, there will be a need for what they refer to as "multiple links" within the systems, which would allow the accessing of data by each group using various levels of aggregation.

As should be clear from the different nature of their responsibilities, there are likely going to be other groupings that the Profitability managers are going to be interested in. Generally, they are probably interested in the functional groupings represented in Fig. 5-4 as Level II. A problem arises, however, because not all of the entities deal with

all of the same products all of the time (a similar problem to that mentioned above). In fact, for some companies, entities which deal with CitiCorp separately (at Level III or II) on some products (perhaps the simple, transaction-oriented products like T-Bills, etc.) may deal as a group (i.e. Level I) with CitiCorp on others. So, the mapping and multiple linking problems are compounded. Therefore, the interface of all of the product level data must have some understanding of all of the actual relationships among the entities of a client and how they differ among various products. This understanding must cover the nature of the relationships as well (i.e. parallel, subsumption, etc.). Further, it must understand how the desired "view" differs across users.

In addition, the interface should probably have an understanding of the way in which aggregated data on one system is to be allocated to a smaller entity level in order for it to be combined with other systems' data for this smaller entity. For example, using the Prudential example, assume I wanted integrated data on exposure to all of Prudential's Insurance entities (Level II). However, I may have data from all products at this entity level except for one: long-term finance. This may be because Prudential found that there were economies to centralizing this function and dealing with bankers as a united front (at Level I). Remember, this is not for Credit management, which only cares about legal entities (and who is ultimately responsible for delivery). The interface can perform two operations in this case: (1) ignore the T-Bill data; (2) allocate it somehow (and perhaps mark it with an asterisk or a footnote). Either way, the interface must have this understanding and processing capability. Clearly, the human beings

that perform this task today have such an understanding. The CIS which might perform it tomorrow must as well.

.2 Instance Identification

The account level identifier (which was discussed above in the section on Credit's integration needs), which remains as probably the single most difficult issue for CitiCorp to manage, is only one type of instance identification problem. There also exists the issue of being able to tie the account to the salesperson and integrating this across products. A chart similar to Fig. 5-3 could just as easily be constructed for the different representations and formats of the salesperson code across the various systems. This instance identification problem, however, is specific to Profitability.

This is further complicated however by the evolution of "client teams". There now exists for those situations the added work of mapping the various codes to the team identifiers in order to be able to generate team-level Profitability figures.

C. Data Integration at Risk Management

This remains an extremely interesting area because they have in fact attempted a certain level of integration at this stage. See Fig. 5-6 for an outline of Utopia as it stands today as well as a description of how most of the components are brought to the interface. The "standardization"

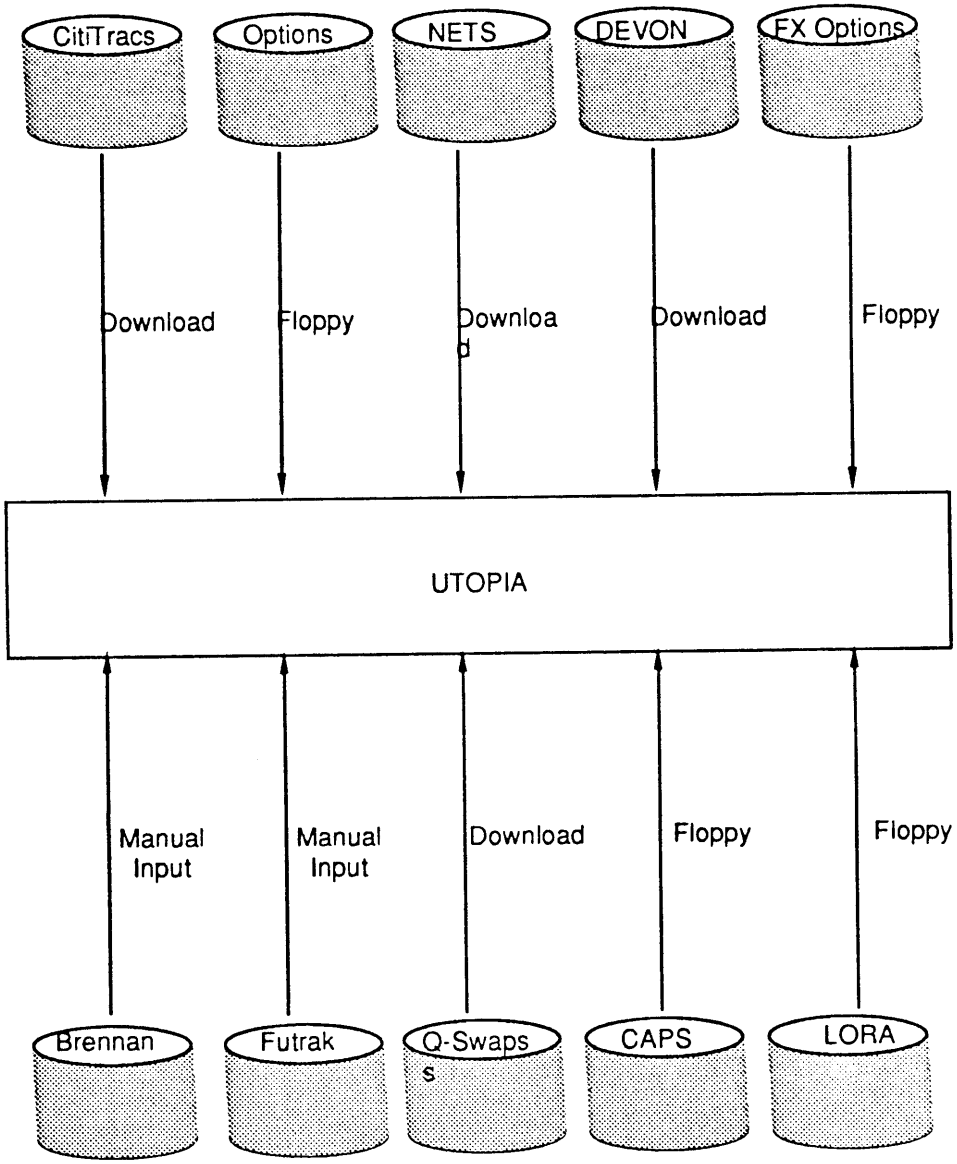


Fig. 5-6: Utopia's Data Sources

process actually is performed locally in each individual system. In other words, there is to some extent a certain level of logical connectivity (i.e. formatting and other problems are taken care of) before there is physical connectivity. Following this section will be a general discussion, using Utopia as an example, of the difference between providing true logical connectivity and simply "downloading and combining."

1. Utopian Evaluation

The designers and operators of Utopia have set a standard format for the information that they will accept for integration. This is different from a typical company-wide attempt at setting technical standards (which would greatly ease the process integration) in that it is an "ex-post" standard which essentially states: "This is how the data should look before it comes into our system, not necessarily as it exists in yours."

While Utopia provides an excellent "integrated environment" for analysis by the Risk manager, it does not provide "integrated information" in the classical logical sense. It is true that the data is generally all in the same format, and it resides in the same PC-based database system, thereby providing a certain level of physical connectivity. However, the data is never logically integrated within the system. The output of the system is a series of product-level screens and/or reports in a common format which surely makes the Risk

managers' analysis function easier, but seems to have far less of an impact on the integration function that they also need to perform. It is this integration that they perform which takes into account all of the interrelationships of the products and their sensitivities as outlined above. The difference between these is to some extent the difference between logical and physical connectivity as shown in Fig. 5-7 (Cases II and III). It should be clear that in this Figure, Case I represents a human finding the data in each separate source, Case II would be much like downloading data into one single place, and Case III is the logical integration of this data (as through a CIS/TK-like system). Utopia's pre-processing would place it somewhere between Cases II and III, yet clearly closer to Case II.

2. Locus of Connectivity

At this point, it is interesting to discuss in more depth the different stages/places at which the standardization (which is not necessarily to say logical connectivity) of the data might take place and the ramifications for each of the location strategies. One could easily draw a parallel between this decision and the design decision for CIS/TK. Where should this particular processing actually take place? The local processing strategy tends to favor data systems that have great differences with few economies to be gained from the centralization of the processing. On the other hand, when there are similarities in the processing that needs to be done, there is likely some advantage to at least some level of centralized processing (such as that which exists in

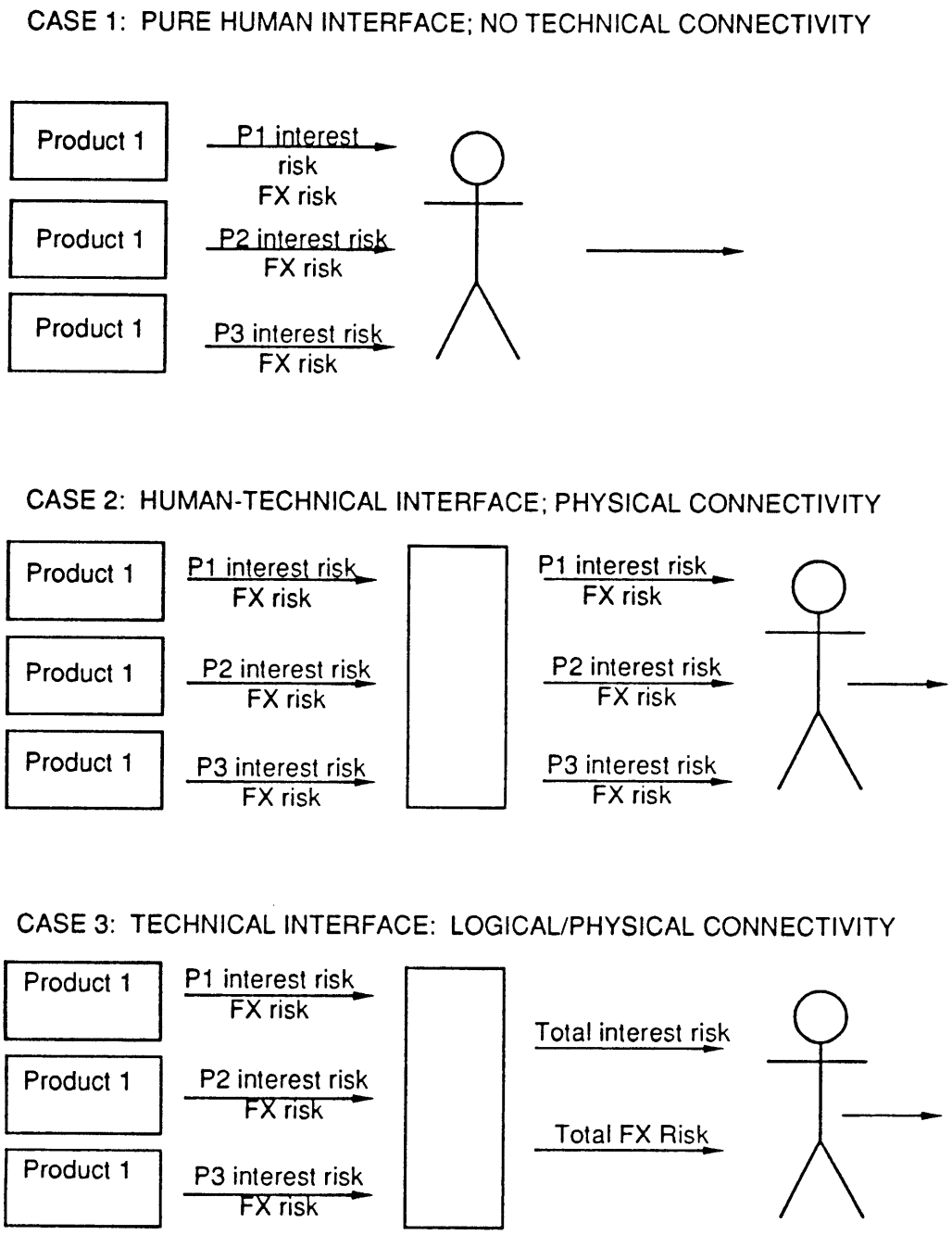


Fig. 5-7: The Three Levels of Connectivity: None, Physical, and Logical

the GQP of CIS/TK). Assume my interface extracted the inventory on-hand at day-end for each product along with its maturity and interest rate and integrated this data for analysis. One day, I might decide that more important than the maturity of the security is its duration (a similar financial measure). Rather than make 20 changes (for each system) in the data set desired as long as the datum is located in the local schema (and there was sufficient mapping) a centralized processor (with the one change) could extract it from all of the systems (as opposed to each of the 20 systems providing it anew). There is clearly a tradeoff between flexibility (which I get with the centralized processor) and complexity (of the design of such a generalized central interface processing system). CIS/TK's design, which allocated responsibilities both locally and centrally, takes what appears to be the prudent middle road approach.

A similar tradeoff existed for the designers of Utopia. It seems likely that the format and content of the data that the Risk Manager processes differ to such a large extent that the added complexity from building a big interface which would "understand" all of the various file formats, etc. was not necessary. Instead, they opted for 20 systems which all understood what they had to output. However, remember that any change that is to be made must be made many times in this structure. The nature of their business clearly values simplicity in this regard over flexibility.

CHAPTER 6: DATA NEEDS OF THE CORPORATE FINANCIAL ANALYST DEPARTMENT (CFAD) OF THE NORTH AMERCIAN FINANCE GROUP (NAFG)

An analysis of the need for, and utilization of, data at the NAFG will provide us with an excellent comparison to those of the NAIB presented in the last chapter. The manifest differences in the natures of the businesses, and the resulting different analytical processes and goals that drive them, present us with a form of data integration far different from that of the NAIB. This chapter will begin by outlining the businesses in which the NAFG competes and what determines success in those businesses. Following that will be a description of the analytical processes that are required to support these businesses and the data that is used in these processes.

I. CitiCorp's North American Financial Group

A. Commercial Banking

Quite different from the business of the NAIB, the NAFG, led by George Davis, is an excellent example of the annuity-type business that was mentioned in Chapter 4 (it may be more accurate to say that this business has traditionally been considered annuity-based). That is, most deals that are structured by this group take the form of a disbursement of cash to the client in return for a promise to pay it back with interest sometime in the future. As mentioned above, the Financial Services Industry has become increasingly competitive in the past decade. As a result, CitiCorp is now not only competing with Chemical Bank, Manufacturers Hanover and Chase Manhattan for customers for their commercial credit services, but they also must sell against the likes of Drexel Burnham Lambert - who underwrite "junk bonds", or non-rated bonds, which are sold publicly - and the borrowing companies themselves who are increasingly apt to bypass the financial intermediaries and issue Commercial Paper directly to the public. Thus, the spreads have thinned due to increasing competition, and more of the income has shifted toward the one time fees associated with the transactions.

To succeed, then, CitiCorp's NAFG must be perform two general functions:

- Identify new marketing opportunities in the form of new borrowers. This might range anywhere from company seeking

cash to buy restructure and buy back its stock to a simple line of credit for a medium-sized manufacturing company.

- Evaluate current opportunities accurately, given the state of the prospect, the expected future macro- and micro-economic environment, and the risk profile of the Bank. This latter factor might cause us to recall the responsibilities of Risk Management at the NAIB. Remember, too little risk may be as bad as too much risk.

Clearly, the analyses performed, and the data that supports these analyses, differ considerably across these functions. As I will discuss below, the role of the CFAD's role is concentrated on the latter, though they do get involved to a limited extent in the former. To sum up the changes of the past decade as they relate to CitiCorp and the bulk of their commercial banking brethren, they must transform their organizations into marketing organizations more than they have ever done in the past, which means shifting some emphasis from financial analysis to market analysis.

The products that they provide and the needs that they fill with these products are at once simple and complex. Each of the products provided by the NFAG, with few exceptions, is the same thing: a debt instrument, and in this sense they are very simple to understand. However, each deal contains a wide variety of different terms (such as period to repayment, indices used for the base price of the loan, restrictive covenants, call or convertible provisions, etc.). This means that, to a

large extent, each deal is different and therefore requires a somewhat different type of analysis.

B. The NAFG Corporate Financial Analyst Department (CFAD)

In 1985 there were in fact no analyst positions at the NAFG. Similar to the situation at the NAIB, however, where CitiCorp initiated an innovative credit risk management process (the Credit Department), market forces presented a problem to the managers at CitiCorp and they responded with an internal change: the creation of the CFAD. The managers of the NAFG noticed that, at that time, there were two roles being performed by the banker: marketing and analysis. It seemed that marketing, which again was becoming of paramount importance, was taking a distant back seat to the constant necessity of performing rigorous analyses.

Now, in retrospect it is easy to see how the different skill sets that each of these functions requires would lead one to the conclusion that they should be separated. However, this has only become true in the past decade or so. Only during this time frame have the marketing requirements become so pronounced in the FSI that such a restructuring would be necessary. In fact, from nearly each person with whom I spoke at CitiCorp I was reminded of the concept of transforming CitiCorp into a "Marketing Organization". It should be made clear that it is this imperative which has resulted in the creation of the CFAD.

The structure of the group is shown in Figure 6-1. Notice the two different segmenting strategies used: geography and industry. The result of this, as will be discussed in more depth below, is that there are many specialized information resources that are used by the specific groups as well as many general resources which the entire analyst pool uses.

Either way, it is a very marketing-oriented approach in that the organization is designed around the user, rather than asking the user to fit their needs within the organizational framework they use. As a contrast, recall the situation at the NAIB. Currently, they are organized around product groupings yet the customer in many occasions is not. Of course, there are several good reasons why this structure exists at the NAIB. However, for marketing reasons, this is why we are seeing the NAIB move toward more of a customer-orientation in their organizational structure (hence the evolution of "customer teams"). Of course, it is just this move toward such a marketing focus which is requiring a higher level of integrated data.

II. Loan Analysis

A. Types of Analysis Performed at the CFAD

Within the second of two major functions performed at the NAFG, financial analysis, there are two different types performed at the CFAD on a routine basis:

- **Annual Review:** This is performed, as its name suggests, each year as long as a loan remains on CitiCorp's books. My impression is that this is a rather routine process that essentially involves the analysis of the company's financial statements (with the limited inclusion of some comparable analyses of competitors and peers). It is used as a monitoring device to ensure the continuance of their ability to repay (as well as several other criteria which will be discussed in the next section) and the compliance with relevant restrictive covenants. The routine nature of the review was so pronounced that there were examples of analysts using Lotus 1-2-3 Macro-based templates which automatically calculated a great deal of the necessary figures (and, further emphasizing the routineness, downloaded the data into a Microsoft Word document complete with a certain level of canned text!).

- **Credit Analysis:** This was described to me as the "sexy" side of the business. It is clearly the more intellectually-stimulating of the two types of analysis and is no doubt what attracts analysts to the position. Credit analysis is performed in order to determine the advisability of the extension of a new debt instrument to a

specific client. As compared with the annual review, it is far less of a science and allows (in fact requires) a high degree of personal creativity on behalf of the analyst in determining which analytical techniques are most applicable to the situation, deciding what data is necessary to support this analysis, and evaluating the results of the analytical methods employed.

The proportion of their time spent on each of the two functions seems to depend on two things: (1) the industry segment (which in turn influences the proportion of new - as opposed to existing - loan agreements); and (2) the job level of the analyst. It seemed that the Senior Analysts (as defined by job title), and particularly the more senior Analysts (defined by experience) were excused from the tedious Annual Reviews which tended to be left for the newer people.

Again, the types of analysis differ mainly in the extent of rote analysis performed. The annual review seemed to be a much more mechanical exercise with little judgement necessary to choose which analytical techniques, which models, and what data were to be used.

B. The Process of a Credit Analysis

Fitting the credit analysis into the framework described in Chapter 1, Fig. 6-2 pictures the general stages of the process. Notice that there are really three different "answers" that the Analyst is looking for. These are collectively known in the commercial banking business as the three "Ways Out" or ways for the bank to recover their investment (with

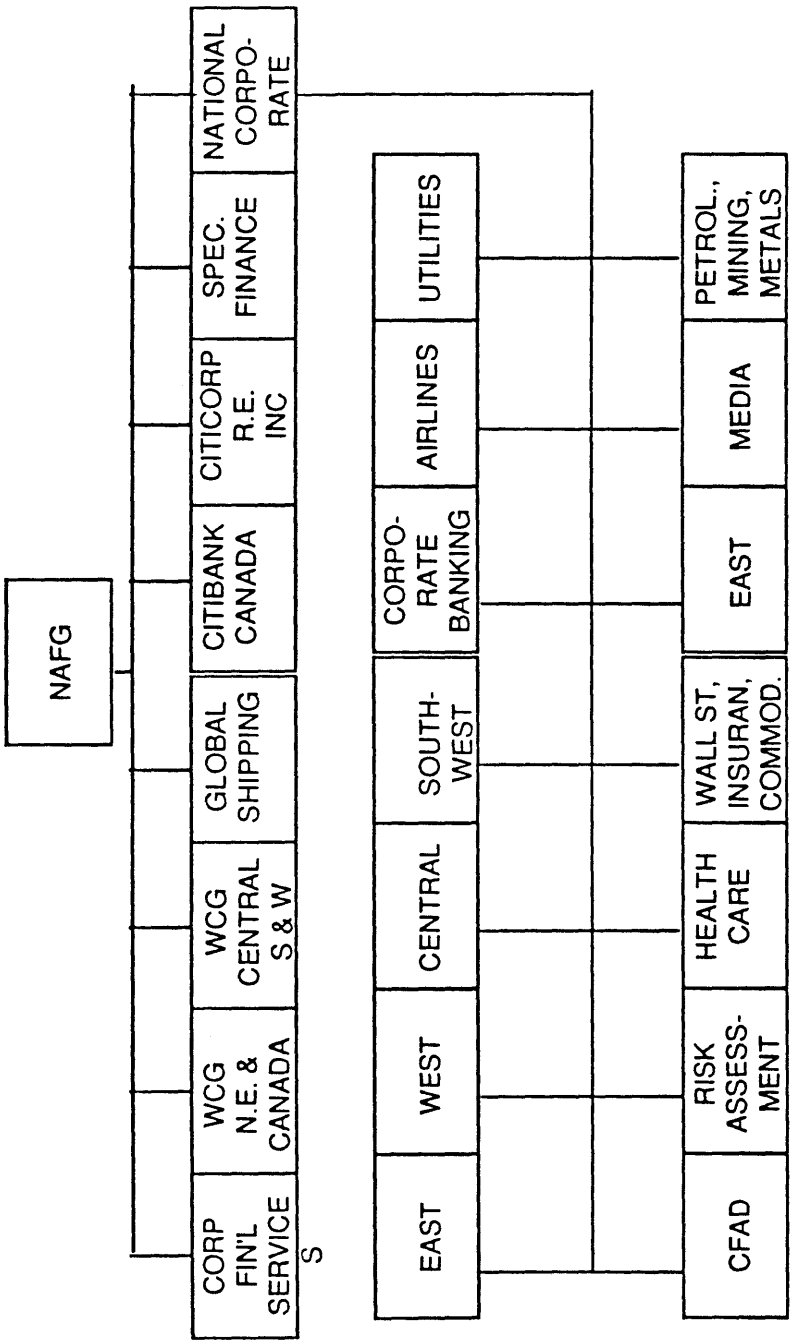


Fig. 6-1: The Organizational Structure of the NAFG

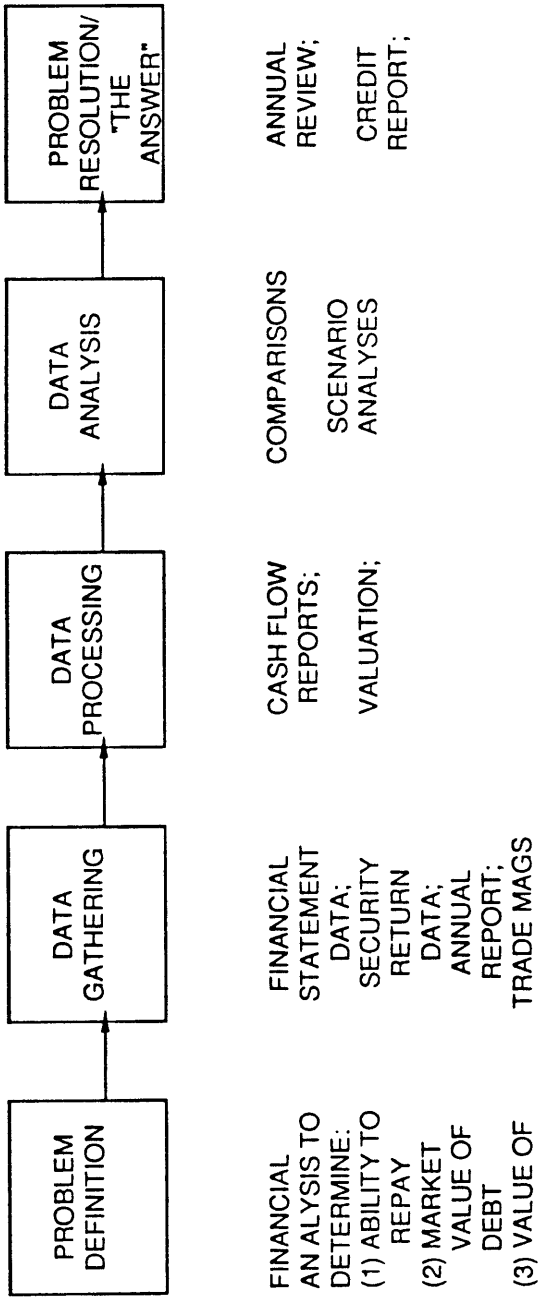


Fig. 6-2: The Nature of Financial Analysis at the CFAD

some level of return, positive or negative). As we progress down the list, we move from more straight-forward analyses to more difficult, abstract analyses and from most-desirable to the least-desirable from CitiCorp's standpoint:

- **Repayment of Loan**: This is the first "Way Out" for CitiCorp, the most common and the most desired from the point of view of the Bank. It involves the payment of all of the principal and interest of the loan as specified by the terms of the contract. The usual method of determining the borrower's ability to do so requires an intensive Cash Flow analysis of the company's operations, using financial statement data as the main input.

- **Capital Markets** : This is the second way that CitiCorp might be able to recover their investment. It involves the public distribution of the debt instrument. The evaluation of this Way Out involves an analysis with a more eclectic approach, investigating the market value of the bonds (including their ratings) as well as the market multiples that they would likely yield in a public distribution.

- **Sale of Assets**: This is generally the last resort for a lender to recover all or part of their capital (unless there is some sort of guarantee, public or private). It involves seizing and selling the real assets of the borrower and/or guarantors. To value the proceeds from such an action, the analyst must ascertain not only their current value but also the future market and macroeconomic trends that might affect that value. Clearly, of all of the analyses

that the analyst must perform, this is the least scientific and the most subject to personal intuition.

C. CFAD Data Sources

The data that the CFAD analyst uses varies from on-line databases to trade journals to rumors they hear on the street (and The Street). The following list of data sources includes the data that seems to be generally used by most of the analysts at some time or another and excludes the specialized sources used by only one or a few groups (which tend to come in hard copy form).

1. Lotus One Source

This is an integrated data product from the Lotus Corporation which provides the user with access to several popular on-line databases. They include:

- **Compustat:** This was explained in Chapter 3. My impression is that up to 70% - 80% (and perhaps more) of the hard financial data that the Analysts use in their formal credit analyses come from one of the several available Compustat databases (which include Industrial, Utilities, and Line of Business (LOB) which includes business segment information using the SIC code as an identifier).

- **Daily Stocks:** A daily time series of stock prices including all types of stocks and warrants. This also includes volume and stock

dividends and splits, etc. It is important to note that unlike those in NAIB, they do not need real-time stock information for their business. This database is updated daily.

• **Bonds**: Similar to the stocks database, it includes daily information about corporate and government bonds. It is also updated daily. One important distinction is that the bond market is not as liquid as the stock market, and therefore the prices that the database provides are IDSI Matrix calculated prices, which are essentially what the price "would have been" had the bond traded. This doesn't seem to be as heavily used as the other databases in One Source.

• **IBES**: Standing for International Brokers Estimate System, this database provides earnings estimates for over 3,400 companies based on the opinions of over 2,500 securities analysts. It is updated each month. This seems to be used fairly heavily.

• **SIC List**: Contains an alphabetically- and numerically-ordered list of industries and their SIC codes. This if for use with other databases.

The Lotus product comes on CD-ROM, but it is downloaded each month in New York and run on a LAN for use by all of the analysts (however certain databases require more timely updating - such as the stock database - which is performed nightly or weekly as prescribed). The processing is all performed locally on CFAD's Banyan system. Most important, access to all of these databases is provided via a "seamless" interface with Lotus 1-2-3. An "add-in" (an extra menu item on the

Lotus menu bar) is provided to allow the analyst easy access to the databases. I will discuss more below about some of the difficulties that an analyst may encounter when using this integrated interface.

2. Other Databases

There are a number of other databases to which the analyst has access. They are listed below (I also got the impression that this is an extremely dynamic list and that new databases may be coming along at any time). While the Systems people claim that Lotus One Source provides up to 75% of the hard data needs of the Analyst, my discussions with the analysts leads me to believe that this number may in fact be larger with the following databases getting only specialized use:

- **M&A Databases:** There are in fact two databases available listing vital historical information on mergers and acquisitions. This is extremely specialized and does not seem to be used very often. Further, while this is apparently not required, access to these databases seems nevertheless to be channeled through one person in the department who "knows how to use them". This makes sense due to the great expense of the connect time and the difficult database navigation protocols. I did not, however, get the feeling that had access been easier and/or cheaper the usage would have been any higher.

• **New Issues:** This includes three databases accessed on-line through IDD which provide data on: Public New Issues, Private New Issues, and Issues in Registration (which have not yet been registered by the SEC).

• **Bank Loan Financings:** They have two on-line databases containing this data which provides such information as the borrower, the purpose, the participating banks, etc. The usage of this was high enough to warrant the planning for network access in the near future (this may in fact have been accomplished by the time of this writing).

• **Bests Insurance Database:** One of the main problems that many analysts noted was the incompatibility of data between the insurance industry and that of other industries as well as the sheer complexity of the business. This database provides such data, but the difficulty in working with it seems to preclude its heavy usage.

3. Quotron

While technically part of CitiCorp, their usage of Quotron appears to be through arms-length transactions. As is always the case with Quotron, the access is made more difficult in that a specialized Quotron terminal must be used. Again, as real-time data is seldom crucial in this business, I was not surprised to see no Quotron terminals on analysts'

desks. However, it seems that this data is in fact used on certain occasions.

4. News Retrieval

In addition to poring over many general and specialized hard copy news sources, the Analysts also seem to make use of electronic news retrieval services. Quotron provides Dow Jones' news retrieval service and Global Report (a product of the Information Business group of CitiCorp) provides their own as well (Comtex). The analysts appeared nearly unanimous in their opinions that Global Report often "missed" a great deal of pertinent news and was therefore of limited value. Global Report does, however, offer the advantage of providing customized templates which allow users to specify a portfolio of companies for which they would like daily news automatically retrieved. This utility, however, does not seem to outweigh the potential cost of missing an important story.

D. Use of Financial Models

At this point, I find it important to define this term as used by CitiCorp. In CitiSpeak, a "model" appears to refer to any automated, or computer-assisted, access to, or manipulation of, data. This includes such seemingly routine functions as printing out canned reports or downloading into a spreadsheet daily stock information about a given portfolio of companies each morning. While I do not doubt the

importance of such utilities, I simply want to distinguish between this and the definition that most people might tend to associate with a "model" which might focus more on sophisticated manipulations of data. Each of the models has been developed within the Lotus 1-2-3 Macro Programming environment.

The following is a brief list of the main models that the Systems group has provided to the Analysts as well as those provided by Lotus:

1. CitiCorp-Developed Models

   • **Comparison Presentation models:** These generally use Compustat to provide historical data in a standard columnar format on the performance of the firm relative to their peers. The analyst must specify the peer group as there is no automatic mapping of peers. In other words, rather than typing in "IBM" and getting a list of relevant indicators for Big Blue and a canned list of peers, I must manually choose, say, Apple, DEC, etc. This probably makes sense given different analyst's opinion as to "peer" relationships. Usually, these models present data in the form of ratio analysis such as ROE, leverage, market/book, etc.

   • **Valuation Models:** These include such models as a a leveraged buyout model, a bond rating model, and a model which determines the market value of the company as the sum of its parts. The data sources of these modes vary, but most tend to use

Compustat in addition to at least one other database such as stocks.

• **Presentation Models:** I have so dubbed these because rather than performing much data-processing they tend to simply present information side-by-side in a spreadsheet (I will discuss the question in chapter 7 of whether this is in fact all that the analyst really wants). Generally, these make use of the various databases available in One Source group.

• **Generic Data Retrieval (GDR):** This model provides access to all of the One Source products and allows the information to be downloaded to the spreadsheet fairly easily. This acts much like a tool for the analysts to develop their own models and appears to be the most heavily used model of those models provided by the CitiCorp systems group.

2. Lotus Models

Lotus also provides several tools for the analysts. Again, to call them "models" may be somewhat confusing. They tend to provide canned report formats drawing from one database (and, less often, from more than one database). These tend to be used only occasionally. The most heavily used model from this group appears to be "MicroScan". This allows users to scan through the databases specifying criteria of companies they would like to select and data they would like to view and/or download.

To summarize, the NAFG needs to perform two functions well to succeed: identify new marketing opportunities and evaluate these and current opportunities. Clearly, anything which would augment their ability to to perform either of these tasks would be of great potential value. In this chapter, I outlined the data that tends to be used at the CFAD in performing the latter function. The next chapter will go into this more deeply within the context of CIS and will look forward to their potential need for a higher degree of data integration.

CHAPTER 7: CONNECTIVITY AT THE CFAD

The previous chapter explained the types of data that are available to the analyst as well as the general types of "problems" to which they are looking for an "answer". It should be clear at this point that they use a very wide array of data sources in their analyses, from on-line sources to hard copy sources to verbal sources. Clearly, the information must all be integrated at some level so that it can all serve to produce the resolution of the "problem". This chapter will discuss the extent to which they are all logically as well as physically connected as well as the very real problems they have run into when attempting either or both types of connectivity within this environment. The focus of the discussion will be on the "knowledge" that must reside at that data interface where the integration takes place. This example is rich in the three possible types of interfaces: human, technical, and mixed.

I. Current State of Connectivity of Data Sources at the CFAD

A. Physical Connectivity

Almost by definition, the totality of data used in the analyses at the CFAD are physically connected in some way. In other words, the information from each of them is brought to the same place for the analysis. However, when discussing connectivity in terms of technical "links", we see that the One Source product essentially represents the bulk of the true technical-interface connectivity. It provides the analysts with a "one-stop-shopping" environment for a great deal of the on-line data that they use on a consistent basis.

This, however, is solely for the One Source products. For the other on-line databases, the analyst is provided with a common machine which serves as a familiar front-end, but he/she must dial them up, and navigate through them himself/herself.

Finally, for the hard copy data, the analysts themselves clearly serve as the interfaces. In fact, there is a specific function performed by one individual at the CFAD to integrate data from a variety of hard copy sources concerning new marketing opportunities (as mentioned earlier). In this role, he is acting as a human interface with the "knowledge" of exactly what constitutes a "marketing opportunity" and where one might look for it.

B. Logical Connectivity

I saw somewhat less connectivity at the logical level. My impression was that this had two main causes:

- **Specificity of Data:** It seems that much of the data sources provided data that was specific to a certain type of analysis. In other words, a cash flow analysis would largely make use of financial statement data (which in this case is provided by Compustat), a bond valuation of bond data. Following all of these analysis of course, the analyst will write this into a report. Therefore, a large proportion of the logical integration tends to be at the textual (and mental) level.

- **Aversion to Over-Automation:** I got a strong sense that most analysts, Junior and Senior, feared any further separation from the actual raw data than already exists. Any automatic integration of data (such as in the complex models which use data from many sources), many of them fear, would reduce the impetus to <u>think</u> about the method of doing so, which might vary greatly depending on the subject involved.

With the combination of these two factors, we see a fairly low level of logical integration of heterogeneous data sources, technical or otherwise, until the final phase of the problem solving process: the analysis of the data. However, there is <u>some</u> such integration as well an obviously high level of intra-database integration. The various problems which result from these will be the focus of the next section.

II. Issues in Connectivity

Again, this example should help greatly in understanding better the specific problems that CIS/TK is being designed to address. I will delineate once more for simplicity the problems related to the physical and those related to the logical level.

A. Physical Connectivity

There were numerous examples of many of the common problems related to physical connectivity that were described in Chapter 2 and demonstrated in Chapters 3 and 5.

1. Modes of Access

As explained above, a great majority of the commonly-used databases were available through One Source and therefore, to some extent, didn't have this problem. However, not all of them were so. Therefore, each analyst, to be able to use as much on-line data as they could for their analysis, needed to understand the protocols for accessing and downloading data from each of the databases.

More interesting, the modes of access within the One Source product line were not completely standard either. While all but one of the databases could be accessed through a standard series of commands using the GDR (Generic Data Retrieval) model, the Stocks database could

not. Its unique format meant that the access to this database was different, to a fairly significant extent, from that of the others. This is likely due to the different nature of a daily updated time series and a quarterly-updated database such as Compustat or IBES.

It was very clear that one could not possibly access any of the most commonly-used databases without a manual at one's side. The variables are coded using a letter to describe the source and a number describing the variable. Therefore, using the Lotus interface, "a1" might be defined as net income by Compustat, while "b6" might represent duration in the bonds database. Clearly, there is rather sophisticated "mapping" (from this applications level to the global level attribute names) performed by the experienced analyst who does not need the manual.

2. Documentation

While the user is provided with a common (and basically friendly) interface, their ability to download all of the data that they want is still limited by their understanding of the databases. I heard from the Systems group several times the following caveat: "The users still need to know what they're working with." Evidently, this means going through all of the different types of documentation and trying to understand exactly what it is that they can get from the system.

In other words, they need to be able to map their needs to the variable names and then to the One Source names in order to get the data. This

means going from "How much have they made each year for the past 5 years?" to "Net Income (Restated)" to "A6". The first step may require a tour through the relevant database's manual while the second could be found either on line in a text file (about which I found few analysts who knew, or cared) or in the One Source manual. There is clearly an initial learning period through which the analyst must progress to get maximum (or perhaps even efficient) usage from the databases.

Over time, it seems that the Analyst finds those data attributes that he/she is interested in and tends to know them, so that this ceases in the long-run (as a result of the learning curve) to become a big issue. However, this obviously limits the analyst's ability to make use of all of the potential sources of data.

3. Lotus 1-2-3 Compatibility

Another manifestation of the level of physical connectivity is the provision of access to all of the (most commonly used) databases through Lotus 1-2-3. This is an environment that they all understand (in fact prefer) since it is there where the bulk of the numerical analysis will be performed.

This, then meant that many of the problems with physical connectivity (such as those that arose at the NAIB where data had to be physically loaded into an integrating system like Utopia) are not really an issue today for the NAFG. However, several of the analysts did express the opinion that for any analysis that you wanted to do which did not fit into the models provided might be rather difficult (to import into the 1-2-3 environment where the other data is), particularly if they wanted

to access more than one database. In other words, they are somewhat limited by the flexibility of the Lotus 1-2-3 environment.

B. Logical Connectivity

As usual , the problems posed by physical connectivity can be solved (as they were at least partially by the Lotus 1-2-3 interface). However, when moving on to logical connectivity, more problems tend to arise that are not quite so easy. This is exactly the case with the CFAD. There are many examples of both explicit and implicit knowledge, a fair proportion of it the latter, that is necessary to acquire the desired data, understand it, and combine it with other data in an insightful analysis. The following is a list of several of the most pressing of these problems. Many of the problems will look the same as those that have been discussed in the previous two case studies. However, they all tend to differ in some way, particularly in the ways in which the users have dealt with the problem.

1. Unique Company Identifier

Recall the similar problem in the previous case studies. In the event study, the formats of the two company identifiers differed with the Compustat cusip being a six-digit number and the CRSP cusip an eight-digit number. This was solved via a fairly straightforward FORTRAN programming device. At the NAIB, the problem was not one of representation, but one of actual differences in identifiers across the

various systems (i.e. not just format, but value). The current proposed solution at the NAIB is one of creating a single unique identifier using the commonly recognized cusip as the basis for the i.d. (of course another problem arose however since non-public companies do not have cusips) and adding other informational numbers to it such as industry, etc.

In the case of the NAFG, the problem is rather different. Each of the databases that they use allows the input of either stock ticker symbols or the cusip to identify the company. Aside from the obvious shortcomings of the cusip, as discussed elsewhere, there is also an important problem with the usage of the stock ticker: it is unique only within a listing stock exchange, not across exchanges. Given this, there is a very real possibility of overlap between the exchanges. My understanding is that the American Stock Exchange (AMEX) and the New York Stock Exchange (NYSE) have very little, if any, overlap. However the problem becomes more acute when working with the Over The Counter (OTC) stocks.

The following example should illuminate this problem fairly well. In 1986, there was an OTC stock offering by a Colorado-based convenient store entrepreneur named R.L. Merchant. The stock ticker assigned to this company on NASDAQ (the tracking system used for OTC stocks) was "RLM". However, on the NYSE, that very same ticker was also assigned to Reynolds Metals, a mining division of R.J. Reynolds. The data retrieval systems used by the analysts (whether its MicroScan or GDR) do not necessarily take into account the exchange on which a company

trades. In fact, the exchange is all but useless information from the standpoint of many of the analyses performed at the CFAD.

As a result, as I was told by one analyst, he was comparing standard numbers on this specific occasion of the metals and mining industry with a convenient store chain. Specifically, he was looking at the R&D numbers, a crucial indicator in the M&M industry but hardly important in a convenient store, and found the numbers to be out of line. This brings up a key point: due to his intelligence, both generally (common sense) and specifically (as an analyst experienced in this industry), he was lucky to have caught this error before any actions were taken based on the erroneous information. He simply knew that a mining company could never spend close to no money in a quarter on research and development and expect to remain a committed competitor in the industry!

So, without the information about on which exchange a company is traded, the interface must have an understanding of the "usual state of affairs" in a given industry, thus being able to spot something that appears out of line as being an error of this type. We will see below knowledge similar to this type (that is, of the "usual state") aiding in the resolution of different problems relating to logical connectivity.

2. Data Integrity

Recall under the discussion of data sources the analysts' opinion on Global Report's news retrieval effectiveness. They didn't like it and

therefore tended not to use it. However, how can a computer judge the integrity of data in situations where the source in general has a good enough reputation to be utilized on a consistent basis? It is clear from my analysis that for a system to serve as an effective interface in this type of domain, such an ability must be built in either through the software itself or through deferrals to the knowledge of the human "expert": the analyst.

It it well known to anybody in a data intensive line of work that there is no such thing as perfect data all of the time. One Source and the other databases that CFAD uses are no different. One example that I came across was a situation in which two companies' data were switched accidently by Compustat (or Lotus). In this specific industry, it seemed that so many of the companies were similar that a cursory look at the balance sheets of many of the players would reveal no significant differences, and therefore would send up no immediate flag that "something is wrong." Due to this generic nature of the industry, the analyst innocently overlooked the error in his preparation of the analysis! In fact, in this case the report went out to the client before one of the Senior Analysts, who was intimately involved with one of the companies and therefore knew that the numbers were obviously in error, noticed the problem and it was eventually corrected. While it was not the case here, it should be clear that a client relationship could easily have been ruined due to this "minor" oversight.

The issue here again becomes the importance of the "human checkpoint". I was warned many times by analysts not to automate too much since a great deal of problems are cleared up in the interim stages

of an analysis by simple "reality-checking" the numbers. This type of checking will be crucial for any interface which intends to send this data to another model (or other manipulator) without first introducing the human checkpoint.

3. Representation of the Data

While I have discussed the problems of different formats and scales previous to this, a couple of different representation issues arose when using Compustat financial data. These were generally intra-database issues. It should be made clear that the precision of Compustat data is extremely important to the analysts due to the precision of their analysis (whereas in my analysis of the market crash, this level of precision was probably not necessary).

The first such issue is the potential for there to be more than one type of format for a given data attribute. This arose when using Compustat data where revenue is generally formatted as a floating-point number (the FORTRAN F8.3) in the millions. However, for approximately ten companies with particularly high sales figures (i.e. IBM, GM, etc.), the format as presented by the One Source-Compustat product actually changes occasionally to an alpha variable with the letter "K" affixed to it. This has obvious implications for the integration of data as well for the manipulation of this data in any model. In other words, were I to build a model which automatically integrated this with other data which expected a number, depending on the environment the alpha might be read as a '0'! This is clearly unacceptable. The interface simply must

have an understanding of when it might receive data in this format and how to deal with it (in this case, to lop off the 'K', convert to a floating-point number and multiply by 1,000).

The other interesting representation problem is the issue of the different methods of presenting financial statement figures (this is another of the many important intra-database issues). Specifically, the differences arose when looking at income figures that were "Restated". This generally takes effect in order to allow users of the data to make more realistic comparisons among firms and attempts to take into account the different accounting methods, etc. However, it doesn't give the Analyst a great picture as to absolutely how the company did (while facilitating the relative measures). An example of the difference in the Earnings Per Share and the Earnings Per Share-Restated values across a sample of companies is shown in columns I and III of Fig. 7-1. Clearly, the values are different enough that any analysis based on a potentially restatable parameter must take into account this potential and evaluate the "pre-processing" that might have been performed on the data within the context of the specific analysis. The differences, again, are generally (yet not always) due to either a merger, acquisition, or a change in the accounting methods used by the Firm.

While there are non-Restated numbers also provided, there tends to be a lesser degree of disaggregation of this data within Compustat, and therefore other sources may have to be used, such as the Annual Report (which will be discussed in more depth below). Finally, while there is some information provided as to how the numbers may have been restated, by most accounts this was not nearly enough information to

Company	(I) Primary EPS	(II) Fully-Diluted EPS	(III) Primary EPS (Restated)	(IV) Fully-Diluted EPS (Res.)
Wshington Homes, Inc.	.36	.31	1.03	.85
Fischbach	-1.010	-1.43	-1.010	2.55
General Cinema	.42	.42	.49	.49
United Merchants	-.76	-.65	-.57	-.57
Texfi Industries	.19	.22	.32	.34
Pope & Talbot	.57	.53	.57	.53
B.F. Goodrich	2.15	2.070	2.3	2.2
Insteel Industries	.39	.35	1.05	.84
Kaisertech Inc	.73	.67	1.06	.95
Union Corp	.19	.18	.3	.29
Fedders	.29	.28	.25	.40
Allegheny Int'l	-.99	-.34	-1.11	-.43
Datapoint	-.18	-.17	-.06	-.06

Fig. 7-1: The Different Calculation Methods of Earnings Per Share

back out the restatements and calculate the true "raw" number. Essentially it informs the analyst that the restatement might be due to acquisitions, accounting changes, etc. but isn't much more specific. However, the cause of the restatement could very well determine the decision by the analyst as to which ("raw", as opposed to restated) to use.

So, any interface which integrates company-level data from Compustat must have an understanding of the method of representation used - such as the level of aggregation, the format, the scale, or the extent to which the data has been changed for any reason - in order to use it effectively in the type of analysis performed in this domain.

4. Method of Calculation

Similar to the situation discussed above with respect to the method of representation, analysts expressed a dire need to be informed of exactly how certain numbers were calculated when they make use of these numbers in some contexts. Again, this time from the analyst side, I was reminded that it is essential to know "what one is working with" (which gives some sort of a hint as to how a number might have been arrived at).

This problem had several different manifestations at different levels of integration:

• **Model Output:** Several of the models, including the valuation-type models, take a company's ticker as a parameter and then go through a certain data manipulation routine and output another number. However, the nature of many analyses of this type is such that each situation is different to some extent (of course there are similarities among many of the analyses performed, however none of the credit analyses are simply routine) and therefore no simple model can give "the answer". On one level, this accounts for the apparent lack of usage of many of the more complex models available. However, it also means that when they are used the analyst must have a good understanding of exactly the method used and how, if at all, the answer must be adjusted for his/her purposes. Of course, it would be unreasonable to expect a different model for every possible permutation involving the multitude of loan terms Therefore, they can't simply plug in numbers and expect the right answer for their analysis. As one particular analyst noted: "an analyst must be wary of a black box."

Any interface which will attempt to provide any level of integration along with some such manipulation or calculation (i.e. a model of some sort) must then make the user aware of the method used or else, using some elements of AI technology, "understand" which method would best be used, or alternatively, understand the ramifications for the rest of the process that the chosen method might have.

Analysis Inputs: In addition to knowing how outputs are calculated, the Analyst must also understand how the inputs are calculated. This is true not only of the inputs to the models discussed above, but also to the analyses that the analysts build themselves. This is another example of an intra-database logical connectivity problem. There are many examples of data attributes with enough subjectivity that they would fall into this category. One such example is the typical ratios that may be provided from Compustat such as Earnings Per Share. As discussed above in Chapter 3, there is a great deal of leeway in reporting this figure, depending on the level of "dilution" (addition of non-equity, but equity-like, securities) one may want to include. Fig. 7-1 shows the differences between primary (columns I and III, only common stock) and fully-diluted (columns II and IV, all common stock equivalents). These differences are significant enough, but there are also many different levels in between that the analyst might be interested in when comparing such levels to other industries. Many other ratios exist with a similar amount of subjectivity.

Another category of attributes which must be fully understood to ensure proper comparison between firms is inventory-based numbers. These include ratios such as Turnover Ratio and Quick Ratio as well as other measures such as Days in Inventory and Cost of Goods Sold. These numbers are subject to a great deal of variability due to the different inventory methods available to the accountant such as LIFO, FIFO, and lower of cost or market. Were

an analyst to simply take any of these numbers as a given, he/she might be place value on a difference (or similarity) among companies that is due to the art of accounting rather than the art of doing business. The simple example in Fig. 7-2 shows how a simple accounting difference can significantly impact the apparent financial strength of a firm. Therefore, any inter-company data integration must take these changes into account (Compustat performs part of this function with their restatements, but as discussed above one can't always be sure as to what is the source of the restatement).

Finally, other accounting differences can have a great deal of impact such as the way in which a company recognizes and categorizes certain transactions. While the difference is most pronounced on an inter-industry basis, it also exists to a rather large extent on an intra-industry basis. These differences might include the method used to recognize revenue (i.e. percent completed method vs. cash method) or perhaps the method used to evaluate accounts receivable (i.e. when to write-off a bad debt, or when to increase a loss reserve). Only by fully understanding these methods can an analyst make true and meaningful comparisons and evaluations.

5. Contradiction of Data Sources:

As with any usage of multiple data sources, CFAD's analyses occasionally run into situations where different sources of the data with the same

Situation:

Purchase two components, one at $10 and then two at $100. Income on the resale of two of these (leaving one in inventory) is $250 each:

LIFO (Last-in, First-out):

Income	$500
COGS	200
Net Income	$300
Inventory	$ 10

FIFO (First-in, First-out):

Income	$500
COGS	120
Net Income	$380
Inventory	$100

Fig. 7-2: Example of The Impact of Accounting Differences on Firm Analysis

meaning (i.e. representing the same thing) yield different values. This seems to mainly occur when comparing Annual Report data (or other client-provided data acquired through the Bank's banking relationship) with information from on-line sources (particularly Compustat). I found that several analysts run regular spot checks of the on-line data against the annual report data. Of course, a great deal of the reason for the contradiction likely is due to the reasons as set out above (different accounting methods for different reporting; restatements; etc.). Fig. 7-3 shows the reality of this problem. While the differences vary, and in this case the sources of the differences are unknown, there must clearly be a level of understanding at the interface as to how to deal with this.

There are several ways in which an interface (in this case the analyst) might solve this contradiction. An average of some sort could be calculated (i.e. when they are close to each other and there is little reason to believe that either is closer than the other to the true figure). Also, an investigation might be launched to ascertain why such a difference occurred and to determine which, if either, was in error (this method might be used if the attribute being measured is of a very precise nature, and therefore any difference implies that one measure may be in error). Alternatively, one of the sources may be assigned a higher "credibility rating" with any contradictions automatically resulting in the adoption of this source's figure. At the CFAD, it seems that most of the time the preferred method is this last one with the data contained in the annual report being the source with the highest credibility. Whether in one of these or some other ways, this exception-handler must be contained in such an interface.

Company	Income per Annual Report	Income per Compustat
Exxon	$5,528.	$5,528.
USX	714	793
General Electric	2,239	2,239
General Motors H	219.2	669.9
IBM	6,582	6,582
Ford Motor Company	2,906.8	2,906.8
General Motors	3,537.2	3,550.9
CitiCorp	1,058	1,058
General Motors E	139.1	323.1

Fig. 7-3: Different Reported Income between Compustat and Annual Report

6. Omitted or non-reported data

Another area in which the Analysts noted a problem when integrating data was the situation in which certain data are not reported for a given company over a given time period. The figure returned by a given data sources varies across all of them from the popular -99.000 to the alpha "N/A" to 00001 (which is used in Compustat). There are two major ramifications of this. To understand the extensiveness of this problem for some data, refer to Fig. 7-4. This shows the percentage of companies reporting the specific Compustat data item over a specific time period.

First, the analysts clearly must understand exactly what the flag is which signals missing data. If this is not understood, the analysis based on the data could be in error. For example, the "N/A" mentioned above might be read as a numeric '0' by some models. This could have dangerous consequences depending on what the model is measuring (as well as on what the actual, unreported, figure might have been).

Also, the user must have an exception-handling routine similar to that for data contradiction that dictates how the flag will be handled. My impression was that most analysts tended to seek alternative sources of data, generally in hard copy form, when faced with such a problem. The technical interface must have such a handler (or else defer to the human user).

Data Item	% Reported in 1988 (for 1986)
Cash	91.0
Inventories	81.5
Current Assets	79.5
Common Equity	98.3
Interest Expense	95.5
Special Items	75.5
Common Outstanding	97.5
Investments and Advances (Equity Method)	69.0
Intangibles	74.0
Labor	25.2
Rental Expense	67.1
Inventory Valuation Method	75.5

Fig. 7-4: % Reported Data in 1988 for 1986 Data

III. Conclusion: Desire for Increased Connectivity

Contrary to the situation at the NAIB, the CFAD's analysts showed less of a sense of urgency for more integration. This makes sense given the nature of the two businesses and their uses of data. At the NAIB, the integration of data will serve to add immediate value to their operation. By integrating, the Credit department will be able to offer a client a single line of credit, Profitability will be able to better monitor the development of their new client teams, and Risk will be able to better control the risks of an increasingly complex environment. These are clear, easily-recognizable (and perhaps quantifiable) benefits to be gained through integration.

The advantages of a higher level of inter-database data integration at the NAFG are slightly more difficult to identify, however. Recall the determinants of success in their business:

- **Identifying New Market Opportunities**
- **Evaluating Clients, Current and Proposed**

Of course, other than the limited news-scanning functions performed in search of new deals, the CFAD is restricted to the latter. This is clearly due to their charter rather than the lack of sufficient data. Therefore the key question is: "How can more integration improve their ability to perform this task?"

At the logical inter-database level, my analysis of their operations revealed little true logical integration (using technical interfaces) of multiple on-line data sources. Most of the hard core financial analysis

seems to be done using Compustat data, which is supplemented by multiple sources of hard copy data. The other analytical methods seem to be done on their own, without too much integration. That is, it seems that the bulk of the analyses are specific to a data source. There are multi-source models available, but their limited applicability is demonstrated to some extent by their lack of extensive use.

On the intra-database level as well as at the physical inter-database level, there is in fact a need and a desire for more integration. In fact, my impression, based on my discussions with the analysts, is that the biggest impact that CIS would have at CFAD specifically (ignoring the CIS impact on marketing ability for now) would be the following:

- Single, friendly interface for access to all data sources using a single command structure.

- Creation of a standard to which new systems would be tailored and therefore "plugged into" to support the single interface.

- "Backtracking" ability which would use AI (or related) technology to back out the various adjustments that the data sources might make, and that might result from different accounting methods, to improve inter-company comparisons. Further, given the domain of the problem (or analyst or group preferences), the system might be able to make further adjustments to even improve such comparisons. For example, an analyst's preference for a valuation analysis might call for a partially-diluted Earnings Per Share figure (according to his/her own dilution formula). This would be an adjustment that CIS/TK

might make in presenting the analyst with the desired, rather than simply available, data (of course, this and other such adjustments would depend on the availability of other necessary data).

- Ensure that all comparison numbers are calculated in the same way (again, perhaps based on preferences at any level within the organization down to the analyst).

- Identify the calculation method. For example, given the EPS and the levels for earnings and various equities, the system could deduce which method was used to calculate that ratio. Again, given a domain (or an analyst preference), it might go a step further and choose the optimal calculation method

- Perform certain "reality checks" which take into account industry and analyzed-company norms. This would attempt to filter out source errors as well as problems such as the instance identification problem which was mentioned earlier. In short, it will ensure some level of face validity.

- The system would clearly need to be accessible from the 1-2-3 environment for reasons as set out above.

Applying this to the CIS/TK environment would yield a picture like Fig. 7-5. The ultimate goal of this application is clearly the presentation to the analyst of the best data possible in terms of "rawness", comparability, validity, consistency, etc. From there, most analysts

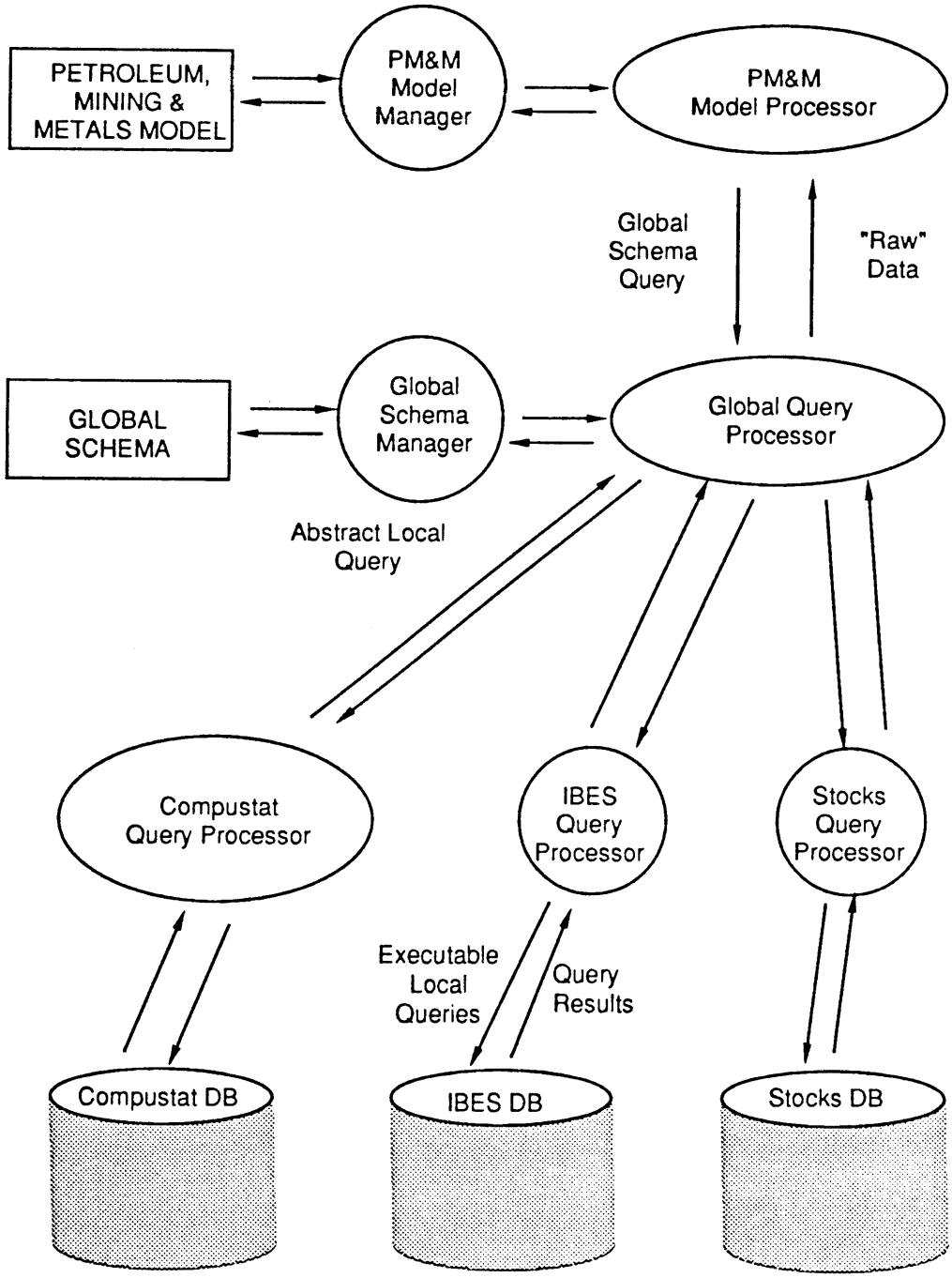


Fig. 7-5: CIS/TK Petroleum. Mining & Metals Financial Analyst Application

seem to prefer to have full control over the manipulations, calculations, and combinations performed.

Not only was such a single interface what the analysts, to a one, mentioned first when asked what it was that they would like. Perhaps more importantly in the long run, it also makes the most sense given the business. Just as the initial impetus for creating the CFAD was to "unbundle" the responsibilities of the banker into analysis and marketing, the system as outlined above would further unbundle the responsibilities of the analyst into data gathering and processing on the one hand and data analysis on the other. Understanding leverage at the CFAD is simple: anything that can help the analysts use their highly-developed analytical skills more effectively is the most certain way of improving the profitability of the CFAD, the NAFG, and (since the NAFG accounted for 1/3 of CitiCorp's after tax profit last year) CitiCorp as a whole. I believe that a system providing functionality as outlined above would do just that.

CHAPTER 8: CONCLUSIONS

This journey has taken us through three examples demonstrating the importance of connectivity in both the academic and the commercial environments. Their diversity was manifest: Example 1 showed how any integrated analysis requires an intelligent data interface. Example 2 demonstrated a similar need with respect to the integration and aggregation of data across dissimilar hardware platforms. Finally, Example 3, while resembling each of the previous two, also demonstrated the very real need for an intelligent interface even when integrating on an intra-database level.

For all of these dissimilarities, the three case studies all had a startling amount of commonality. To conclude this thesis, I would like to present some of these common themes and what they mean to both the CIS developer and to the user:

- **Composite Information Systems are Inevitable**: The ability to integrate data is becoming crucial in the commercial environment. The driving forces behind this trend include: stiffening global competition, the proliferation of on-line data sources, and a trend toward end-user-based marketing (as opposed to product-group-based marketing). An example of this latter trend is the Financial Services Industry.

• **Building a CIS is Not Easy**: While this may seem like an obvious statement, it is ever so crucial. The difficulty one has in clearing the "first-order issues" is often child's play when compared with the monstrous task of facing the "second-order issues" which stand between the user and real logical connectivity. The case studies offer rich examples of the nature of each of these groups of issues.

• **A CIS Must Be Dynamic**: The rapidly changing environment in which these systems will be implemented dictate that a CIS must be an organic system. It must be flexible enough to take on new data sources as well as provide new forms of logical integration as the necessity arises. The former case was most pronounced at the NAIB, while the latter would be crucial at the NAFG.

• **Many Similar Problems** existed in the integration efforts of each of the different examples. The instance identification problem, formatting, and scaling issues were all present, though in various forms, in each case study.

• **Customization vs. Standardization** would be a key issue in each case. This is clearly expected given the nature of the CIS: an amalgam of many different sources of data, which are likely used by different people in different ways. At the NAIB, for example, each of the three user groups, while using the same data (a standardized global schema), used it in very different ways. Further, at the CFAD, we saw how each analyst, while generally

drawing data from the same databases, performed different forms of analysis, depending on the nature of the specific transaction, on the specific focus of the group, or perhaps on personal preference.

Drawing this all together, it should become clear that a system architecture similar to that outlined in Fig. 2-1 (CIS/TK) would be a very effective means to meet these criteria in building a CIS. Clearly, the LQP allows the system to accept new data sources fairly easily. This would be essential in a case such as the NAIB where new products are being developed constantly (and old ones phased out).

Rather than re-invent the wheel whenever a new data source is brought on line, however, the GQP provides a certain level of standardization as it performs much of the processing common to all of the users. It also provides the standard (or global) schema which acts as the "master database" for all of the applications. In so doing, the GQP provides a potential for economies of scale in building a CIS, given a certain proportion of "common problems", which as argued throughout the thesis most certainly exists.

Finally, the crucial ability to customize the interaction between the user and the system is made available through the AQP. Recall the CFAD financial analyst again. The ability to provide the analyst with data that he/she prefers (i.e. in whatever state of "rawness" or "restatement") would represent a significant advance over the current state of information systems in such an environment.

In summary, CIS/TK is clearly facing the "right" problem in the sense that it is a very real one, and only continues to become more pronounced. The successful implementation of this or any other similar system will depend not only on the recognition of the needs of the users and the benefits that they seek from such systems. It will also require the correct classification of these needs (general, specific, etc.) and translation into the functionality offered by the various components of the system. It is hoped that these three case studies will provide a certain level of direction for both the recognition and the classification of these needs.

Bibliography

Frank, Madnick and Wang [1987], "A Conceptual Model for Integrated Autonomous Processing: An International Bank's Experience with Large Databases," Proceedings of the 8th Annual International Conference on Information Systems (ICIS), December 1987, pp/ 219-231.

Henderson, Rockart, and Sifonis[1984] Henderson, John C.; Rockart, John F.; and Sifonis, John G.; "A Planning Methodology for Integrating Management Support Systems," Center for Information Systems Research Working Paper No. 116, Sloan School of Management, MIT, Cambridge, MA, September, 1984

Henderson and Sifonis[1988] Henderson, John C. and Sifonis, John G., "The Value of Strategic IS Planning: Consistency, Validity, and IS Markets," MIS Quarterly, June 1988

Parsons [1983] Parsons, Gregory L., "Information Technology: A New Competitive Weapon," Harvard Business Review, April 1985

Porter[1979] Porter, Michael, "How Competitive Forces Shape Strategy," HBR, April 1979

Appendix A: List of CitiCorp Contacts

Name	Group	Area
Judy Pessin	NAIB	Operations/ Systems
Dorothy Conroy	NAIB	Systems
Evan Picoult	NAIB	Risk
John Remmert	NAIB	Profitability
Bud Berro	NAIB	Credit
Tracey Peter	CFAD	Analyst
David Lipfert	CFAD	"Information Consultant"
Stephen Ellis	NAFG	Division Exec./ Systems
David Moore	CFAD	Analyst
Ken Weinstein	CFAD	Senior Analyst
Jay Newbury	CFAD	Senior Analyst
Gary Geresi	NAFG	Systems
Rita Terdiman	NAFG	Business Solutions