Evolution to Very MANY Large Data Bases: Dealing with Large-Scale Semantic Heterogeneity

Stuart E. Madnick

Sloan WP# 3847 CISL WP# 95-03
August 1995

The Sloan School of Management
Massachusetts Institute of Technology
Cambridge, MA 02139

From VLDB to VMLDB (Very MANY Large Data Bases): Dealing with Large-Scale Semantic Heterogeneity

Stuart E. Madnick
Sloan School of Management, Massachusetts Institute of Technology
Cambridge, Massachusetts 02139 USA
[smadnick@mit.edu]

ABSTRACT

The popularity of distributed computing environments and the growth of the "Information SuperHighway" have dramatically increased the number of data bases available for use. Unfortunately, there are significant challenges to be overcome.

One particular problem is *context interchange*, whereby each source of information and potential receiver of that information may operate with a different *context*, leading to large-scale semantic heterogeneity. A context is the collection of implicit assumptions about the *context definition* (i.e., meaning) and *context characteristics* (i.e., quality) of the information. This paper describes various forms of context challenges and examples of potential *context mediation services*, such as data semantics acquisition, data quality attributes, and evolving semantics and quality, that can mitigate the problem.

1. INTRODUCTION

Advances in technology resulting in increased storage capacity and reduced cost, combined with the needs to gather and analyze enormous amounts of data have lead to the development of Very Large Data Bases (VLDB's.) More recently, the popularity of distributed computing environments (often termed "client/server" computing) have produced an environment that supports and even encourages the development of increasing numbers of databases in organizations. Furthermore, the growth of the Internet "Information SuperHighway" offers the possibility to access information from around the world in support of many important applications in areas such as finance, manufacturing, and transportation (e.g., global risk management, integrated supply chain management, global in-transit visibility.) Thus, we are entering an era of Very MANY Large Data Bases (VMLDB).

The "Information SuperHighway", and its current form as the Internet, has received considerable attention in government, business, academic, and media circles. Although a major point of interest has focused on the rapidly increasing number of users, sometimes estimated as 20 million and growing, an even more important issue is the millions of information resources that are becoming accessible.

Today, when people talk about "surfing the 'net," they usually refer to use of the World Wide Web (WWW) through some user friendly interface, such as Mosaic or Netscape. This type of activity can be effective for casual usage but requires significant human intervention for navigation (i.e., locating the appropriate sources) and interpretation (i.e., reading and understanding the information found.)

Consider the opportunities and challenges posed by exploiting these global information resources in an integrated manner. Let us assume that we have access to information from each of the various stock exchanges (possibly with a delayed transmission for regulatory purposes) and each of the weather services around the world. We might want to know the current value of our international investments, which might require access to multiple exchanges both in the USA (e.g., NYSE, NASDAQ) and overseas (London, Tokyo). As another example, you might want to know where are the best ski conditions at resorts around the world. To manually access and interpret the numerous information sources relevant to these examples would rapidly become impractical. Although some problems may be immediately obvious, there are subtle but important challenges also.

A major such challenge is *context interchange*, whereby each source of information and potential receiver of that information may operate with a different *context*. A context is the collection of implicit assumptions about the *context definition* (i.e., meaning) and *context characteristics* (i.e., quality) of the information. When the information moves from one context to another, it may be misinterpreted (e.g., sender expressed the price in French francs, receiver assumed that it meant US dollars.)

This paper describes various forms of context challenges and examples of potential *context mediation services*, such as data semantics acquisition, data quality attributes, and evolving semantics and quality, that can mitigate the problem.

Proceedings of the 21st VLDB Conference
Zurich, Switzerland 1995

2. THE ROLE OF CONTEXT

Increased information integration is important to business in order to improve *inter-organizational relationships*, increase the effectiveness of *intra-organization coordination*, and provide for much more *organizational adaptability*. Examples of these opportunities and their importance can be found in *The Corporation of the 1990s: Information Technology and Organizational Transformation* [10].

There is an important concept which we will refer to as *context*. In order for people to use information, whether electronic or other media, there is a need for context, which is the way in which we interpret the information. That is, what does it mean (which we call the *context definition*) and how good is it (which we call the *context charactersitics*.)

The context is not explicit for at least two reasons. First, it provides efficiency of communication (e.g., if asked what is your grade point average by a fellow student, you can reply "3.8" without having to explain that it is a 4-point scale with 4 being best, etc.). Second, the context can be so fundamental in an environment that most are not even aware that there is another possible interpretation (e.g., we all know that a grade of "A" is 4.0.)

2.1 Context Differences

Context may vary in three major ways. First, context varies due to *geographical* differences, that is, the ways things are interpreted in the US is different from that in England, France, or China. Second, there are *functional* differences. Even within the same organization and location, different functional areas interpret and use information differently. Third, there are *organizational* differences. The information used in the same function, in the same industry, in the same country, can have different meanings between two companies. For example, the way in which CitiBank might define a credit rating could be different from the way Chase does the similar thing. Thus, context can differ from one organization to another.

Previously, people, information, and context were tightly coupled. For those in charge of cash management in a financial organization in New York City, the fact that they deal with the world in a particular way is not a problem because the information used and the people who use it are all together in one place and share the same context. In that same city a different function of the organization, loans for example, may operate differently but independently. Further, the same activity, such as cash management in New York City may or may not be identical to the same activity in London. The point is that although these contexts can differ, as long as the people, information, and context of a group all remain coupled together and separate from all other groups of people, information and their contexts, there is no problem.

The business needs to integrate information and the advances in technology that make it physically possible have combined to produce both good news and bad news. The good news is that now we can communicate electronically in seconds or fractions of seconds, gathering information from many data bases throughout our organization or from related organizations all over the world. The trouble is that we can gather the information, but the context gets left behind. We can ask for the price of an item and get an answer such as "23", but is that $ or £? Is it single $'s or thousands (as an aside, even if given a clue, such as 23M, there may be a problem because sometimes M means millions, sometimes it means thousands -- in which case MM is used to mean millions)? Is it for a single item or a group (e.g., block of shares)? Does it include or exclude taxes, commissions, etc.? The answers to these questions are usually well known to the traditional users of that source information and that share its context. In financial organizations, for example, this situation creates great problems in areas such as risk management, profitability analysis, and credit management where information must be gathered from many sources with differing contexts. In order to be effective all these applications require information from many data bases, but it must be integrated intelligently.

2.2 Challenges in a VMLDB Environment

In a VMLDB environment, the above examples represent serious problems. Information gathered from throughout the world in different organizations and different functions has many individual contexts, contexts that are lost when the information is transmitted.

Although some may think the solution is to come up with a single context for the whole world, or at least all of the parts of the same company, in reality this is extremely difficult for any complex organization. There are often real reasons why different people, different societies, different countries, different functions, different organizations may look at the same picture and see something very different [16]. To assume that this can be prevented is a mistake. We must accept the fact that there is diversity in the world, yet we need to integrate information. The challenge is to integrate global information from diverse organizations but to take the context differences into consideration. This paper focuses on how technology can help us to meet this challenge.

3. EXAMPLE APPLICATION

3.1 Component Systems

Let us consider an actual situation with only two information sources. One is a service of I.P. Sharp, called Disclosure, from Toronto, Canada. This service provides financial information on companies such as their profits, sales, number of employees, etc. This system focuses mainly on North American companies. The other service, operated by Finsbury Data Services out of London, called Dataline, has information primarily on European companies, their sales, profits, and number of employees, etc.

3.2 Context Challenges in the Example

Some of the typical problems encountered are illustrated in Figure 1. The information on the left-hand side emanates from the Disclosure system, and the information on the right from the Dataline system. Both of these systems have information on the HONDA automotive company.

DISCLOSURE	
ATTRIBUTE	VALUE
COMPNO	3842
CF	19,860,228
NI	146,502
NS	2,909,574
NRCEX *(ROE)*	0.11

	DATALINE
VALUE	ATTRIBUTE
HOND	CODE
28-02-86	PERIOD END
146,502	EARNED FOR ORDINARY
2,909,574	TOTAL SALES
19.57	RETURN ON SHAREHOLDER EQUITY

Figure 1: Context Differences in Information Sources (Information on HONDA from Disclosure and Dataline)

Identification differences. First, for rapid access to the information in Disclosure you would need to know the company's COMPNO, which is 3842, whereas in Dataline you would need to know its CODE, which is HOND. Assuming that you were able to get the above information, let us see how the rest of the information shown in Figure 1 can be interpreted.

Format differences. Note the "period ending" information for the Dataline system, it is 28-02-86. Notice the order in which this date is indicated, with the day first. If you are American you would say that the day and month are backwards because it is our custom first to represent the month, then the day. This problem can be recognized because there is no month 28. In only seven or eight years there will be dates like 01/02/03! What is this? Is it January 2, 2003? Is it February 1, 2003? Is it February 3, 2001? If this is the shipping date, it can make a big difference.

This problem with dates is quite common and some systems attempt to solve this specific problem by means of predefined data types for dates, such as mm-dd-yy and dd-mm-yy. Unfortunately the variety of potential format differences usually exceeds the foresight of predefined data types. In the example of Figure 1, what date does 19,860,228 correspond to? It is really 1986-02-28, that is February 28, 1986. In this system all data, including the date, are displayed as financial data, with commas and a dollar sign in front. (Which lead one of my students to remark, "this proves that time really is money!") For those who use this system every day, there is no problem -- it is obvious to them. But for those who do not use it every day, clearly it is confusing.

Attribute naming differences. As some of the above examples also illustrate, the attribute names, such as CF, are not necessarily obvious since abbreviations are often used. Thus, NI for Net Income and NS for Net Sales are typical and may be decipherable whereas NRCEX for Return on Shareholder Equity may not be quite so obvious. The attribute named "Earned for Ordinary" requires a bit of context background to understand. What are called "common shares" in the USA, are called "ordinary shares" in the UK. In typical UK accounting reports, the profits or earnings of the corporation are referred to as "Earnings attributable to the Ordinary Shareholders" or "Earned for Ordinary" for short. A reasonable abbreviation for a UK accountant but likely to be a puzzlement to a USA analyst.

Scale differences. At the bottom of this table there are two numbers. These are return on equity. These numbers illustrate a scale difference. On the left-hand side, the number is expressed as a decimal fraction -- 0.11. On the right-hand side, it is expressed as a percentage -- 19.5%. That is a difference that could have a significant impact if not realized.

Definitional differences. What is more fundamental, and more puzzling, about this return on equity example is that one number is approximately 11% and the other is approximately 20%. How can the same company in the same year have two different "return on equity" values that vary by a factor of two? And yet there is no mistake, it is not a typographical error.

Anyone with accounting experience would know that return on equity is return divided by equity. However, this opens the question of what is meant by "return" and "equity"? Within generally accepted accounting principles there are many variations of interpretation (e.g., how are extraordinary expenses handled, what depreciation rules are used, how are certain types of stock options handled?). Starting with the exact same raw data, the Disclosure people came up with one number, and the Dataline people came up with a different number. Both are correct -- for their own context.

When we tried to determine what does 'return on equity' mean in each database, we encountered considerable difficulty. Multiple steps were required: the local customer support people normally deal with simple questions, such as "what does this command do" or "what is the charge for your service?", so we were referred to their local data expert -- who did not know the answer. Then the search moved to their headquarters support staff and likewise went from the first-line staff to the data experts. One unexpected problem that increased the difficulty and time effort is that one of the companies got some of the data from another company and resold it, so they did not even know how it was calculated or what it meant, they just got it and passed it on. The key point is that as information goes from organization to organization, flowing around the world, we have more and more information but we know less and less what it means.

Inter-database Instance Identification. There is another frequent problem not shown in Figure 1, which we refer to as "inter-database instance identification." In MIT's Alumni database you would find a company called "Ford Motor Co." In MIT's Placement database there is something called "The Ford Motor Company." In Disclosure there is something called "Ford Motor Co", and in the Dataline database there is something called "Ford Motor USA."

In short, for the same company there are four different ways that the name was recorded. At first it might look like at least two are the same, but there is a subtle difference. The first has a period at the end of "Co." for Company whereas the third does not have the period. We might not have noticed this, but computers would have viewed them as different names if we tried to do a data base join.

4. INTEGRATION CHALLENGES

Although the example of the previous section was simple, such needs and problems occur throughout all businesses. Other examples of actual situations can be found in [3, 9].

There has been tremendous successes within local systems -- the systems that do sales; the systems that do inventory; the systems that do forecasting within the autonomous parts of our organization. The challenge is how to tie these systems in with other functions in the organization, with other geographical parts of the organization, and with partners: suppliers, customers, and other forms of allies.

These types of problems have existed for a long time. Traditionally they have been solved by determining the translations needed and either performing these translation by hand or by writing custom programs. These translation may be directly system to system or via one or more global schemas. As long as the number of data bases involved was small and their contexts fairly constant, this was a viable strategy. But, as the number of data bases continues to increase dramatically with new ones constantly being added and there is increasing volatility in the contexts of the sources and receivers, this manual approach becomes infeasible and new strategies must be developed.

5. CONTEXT MEDIATION SERVICES

Effectively integrating information from multiple sources both within and across organizations represents an important solution to many critical business needs [4], but a key challenge for integration technology research [7, 11]. Organizations can be simultaneously "data rich" and "information poor" if they do not know how to identify, categorize, summarize, and organize the data. Although there are many important integration technology research directions in this area, three particular examples will be highlighted: data semantics acquisition and conflict resolution, data quality, and data semantics and quality evolution. We refer to these types of efforts as *context mediation services.*

5.1 Data semantics acquisition and conflict resolution

As business operations become increasingly dispersed geographically and functionally, differences in work processes at each site performed by people trained for each site lead to data incompatibilities and inconsistencies when these differing sites must interact. Before these differences could be reconciled, we would need to be able to represent the semantics of the data as used in each environment, what we have called the *context* of the data [14]. Research on using metadata to represent *context definitions* provides the basis for capturing and disseminating knowledge about data meanings and can facilitate the data reconciliation and integration process [12, 13]. In this particular approach, as illustrated in Figure 2, the context of the sources (called the *export context*) and the receivers (called the *import context*) is captured. A receiver may be a human, an application, or another database. A *context mediator* then compares the contexts and determine if they are same; if they are not, it attempts to translate the source information into the receiver's context using general *context conversion knowledge* (e.g., it knows how to convert from feet to yards, $ to £, without-tax to with-tax).

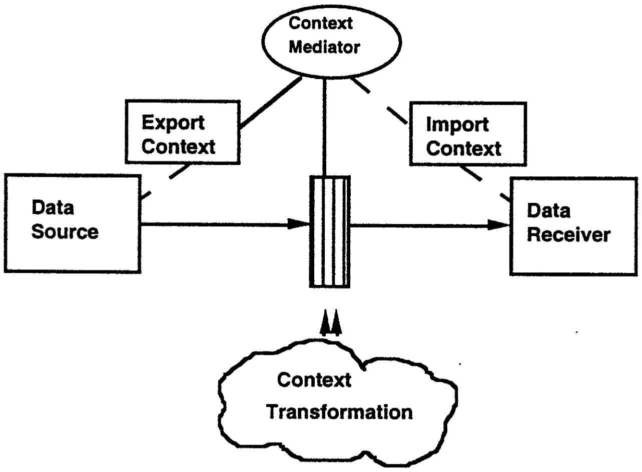

Figure 2. Context Interchange Architecture

5.2 Data quality

Organizations have become very concerned about quality in areas ranging from manufacturing quality to software program quality. Data quality, in comparison, has received relatively little attention. Issues relating to data quality are becoming increasingly important as information is moved through multiple organizations. To a large extent, data quality considerations in the past were handled through personal familiarity; the user knew the characteristics of the data used in his or her organization and informally took this into account when using the data. This approach is not feasible as increasing numbers of information sources are used, many not well known to the user. We are increasingly exposed to data with various levels of quality for which we do not have first-hand familiarity. Furthermore, many currently automated processes for converting, merging, and manipulating the data renders inaccessible information about the original data that might have conveyed information about its quality. For example, the originating source of a given piece of information is often a key element in judgements about its credibility and quality [8, 18].

There are many data quality attributes that may be important, such as accuracy, completeness, consistency, timeliness, source, and stability [15, 17]. Defining and measuring the important data quality attributes, which we refer to as *context characteristics*, is an important step. Even a simple concept as "accuracy" can have many subleties. For example, in the July 21, 1995 issue of the *Wall Street Journal*, an article entitled "Nasdaq Pushed Past Capacity in Latest Surge" described the situation of Henry Cohen who wanted to check the Nasdaq Stock Market composite index and "searched four different places -- three newspapers and his mutual-fund company -- and got four different closes or changes in the index's value." None of these were typographical errors, each was at some point in time the best and most accurate answer. But due to the market's tumult that day, there were many subsequent revisions to the index's calculations for that day. Thus, accuracy in this case was time-dependent.

It is necessary to properly maintain this quality-related information as data moves through and between systems, as part of the context mediation services. The defining, measuring and propagation of context characteristics represent significant challenges and important research areas. But, with this quality information, decision makers would be better able to make effective use of the data.

5.3 Evolving semantics and quality

It must be realized that autonomous databases are independently evolving in semantics and quality as well as in content (i.e., values). For example, consider the situation of stock exchanges around the world. Not only are the stock prices changing continuously, but the definition of the stock price also can change. At some time in the future, the Paris stock exchange may change from being measured in French francs to ECUs (European Currency Units). The normal "ticker tape" data feeds do not explicitly report the currency, it is implicit in the context of the source.

Although the example of changing francs to ECUs is currently hypothetical, last year the Nasdaq (the USA over-the-counter stock exchange) changed to reporting prices in units of 64ths (code #) in addition to reporting in 16ths (code *) and 32ths (code /). This change caused enough problems to have received front page coverage in the *Wall Street Journal*.

More subtle examples include changes from reporting "latest nominal price" to "latest closing price" or from a percentage based pricing to actual prices, as happened at the Madrid stock exchange. Furthermore, in a historical database of stock prices, it must be recognized that the meanings had changed over time especially when doing a longitudinal analysis.

Of course, the quality of the sources and the quality requirements of the receivers also change over time. In many cases, new data capture technologies and procedures can improve the quality. Alternatively, cost-cutting measures or organizational and procedural changes may decrease the quality. Likewise, the receiver may need higher quality information due to its more critical role in decision-making or may be able to settle for lower quality due to its less critical role or the available of additional comparison sources.

By capturing the context of the sources and receivers, the *semantic context mediator* can formally and automatically compare the contexts to determine if they are compatible, partially compatible, convertible, or incomparable. Similarly, by representing the quality characteristics of the source data and the quality needs and tradeoffs of the receiver, the *data quality context mediator* can determine if they are compatible. These mediator services can be performed on a dynamic basis. Thus, as source or receiver contexts change, the necessary adjustments are made automatically allowing the autonomous evolution of the individual systems. This is a critical requirement in most environments and an important premise for the growth of the Information SuperHighway and the emergence of Very Many Large Data Bases (VMLDB). The research efforts on *context knowledge* [13] and our Context Interchange Prototype system [1] represent directions towards solving the more general problem of *context interchange* [2, 14].

6. CONCLUDING COMMENTS

A key challenge in effectively integrating global information and exploiting the capabilities of the Information SuperHighway is our ability to tie the contexts together. There are systems now being developed to deal with this challenge, which over the next few years will rise in importance.

One dramatic example of the importance of these efforts can be found in a US government study of lessons learned during the Gulf War. There was a tremendous transportation coordination issue involved since over 70% of all materials shipped to the Gulf used commercial carriers: commercial trucks, trains, ships, planes, with material coming from diverse commercial companies -- food companies, clothing companies and so on.

According to this study, there were about thirty-thousand huge containers of material shipped from around the world -- much from the US but also from elsewhere -- to the Gulf theater of operation. An occasional container would arrive at the Gulf with no information to explain what was inside. These containers then had to be opened and all the materials unloaded and inventoried at a port in Kuwait or Saudi Arabia, repacked once its contents were identified, and then shipped on to an appropriate destination. Of the thirty-thousand containers shipped to the Gulf over 27,000 containers had to be hand-inspected.

The point here is that *we can move containers weighing tons around the world faster than we can move the needed information to tell us what is in these containers*. This is not an issue of "stupidity" or human error. It results because there are hundreds of different computer systems and data bases in the airlines, the shipping companies, the port facilities, railroads, trucking companies, and manufacturing companies. These systems were never designed to directly operate with each other. Although each system may be efficient, the interfaces between these systems, effectively the "on and off-ramps" of the information highway, introduce tremendous disruptions and delay, usually necessitating significant human intervention or specialized handling. This situation is not limited to the US military since the same requirements exist for all large organizations.

In conclusion, there is a fantastic opportunity to economically and efficiently capture and store enormous amounts of information in Very Many Large Data Bases. But there is a critical need to deal with large-scale semantic heterogeneity if we are to be truly effective in integrating such systems. Exciting opportunities and challenges lie ahead for all of us.

Acknowledgements

Advice and feedback from Yang Lee, Diane Strong, Veda Storey, Richard Wang, Michael Siegel, and Amar Gupta are gratefully acknowledged.

Work reported herein has been supported, in part, by the Advanced Research Projects Agency (ARPA) and the USAF/Rome Laboratory under contract F30602-93-C-0160, the MIT International Financial Services Research Center (IFSRC), the MIT Leaders For Manufacturing (LFM) Program, the MIT PROductivity From Information Technology (PROFIT) Program, and the MIT Total Data Quality Management (TDQM) Program. Information about the Context Interchange project can be obtained at http://rombutan.mit.edu/context.html.

REFERENCES

[1] A. Daruwala, C. Goh, S. Hofmeister, K. Hussein, S. Madnick, and M. Siegel, "The Context Interchange Network Prototype," to appear in the *Proceedings of the Sixth IFIP TC-2 Conference on Data Semantic (DS-6)*, 1995.

[2] C. Goh, S. Madnick, and M. Siegel, "Context Interchange: Overcoming the Challenges of Large-Scale Interoperable Database Systems in a Dynamic Environment", *Proceedings of the Third International Conference on Information and Knowledge Management (CIKM 94)*, November 1994.

[3] R. L. Kay, "What's the Meaning of This?", *Computerworld*, October 17, 1994, pp. 89-93.

[4] S. Madnick, "Chapter 2: The Information Technology Platform", in *The Corporation of the 1990s: Information Technology and Organizational Transformation*, M. S. Scott-Morton (Editor), Oxford University Press, 1991.

[5] S. Madnick, "The Challenge: To Be Part of the Solution Instead of Being the Problem", *Proceedings of the Workshop on Information Technology and Systems* (WITS'92), December 1992, Dallas, Texas.

[6] S. Madnick, "Chapter 16: Putting IT All Together Before it Falls Apart", in *Information Technology in Action: Trends and Perspectives*, Richard Y. Wang (Editor), Prentice-Hall, 1993.

[7] S. Madnick, M. Siegel, and R. Wang, "The Composite Information Systems Laboratory (CISL) Project at MIT, *IEEE Data Engineering*, June 1990, pp. 10-15.

[8] S. Madnick and R. Wang, "Introduction to the TDQM Research Program", TDQM Report TDQM-92-01, MIT Sloan School of Management, Cambridge, MA, May 1992.

[9] P. Quiddington, "Cruising Along the Information Highway", *MIT Management Magazine*, Fall 1991.

[10] M.S. Scott-Morton, (Editor), *The Corporation of the 1990s: Information Technology and Organizational Transformation*, Oxford University Press, 1991.

[11] M. Siegel, S. Madnick et al, "CISL: Composing Answers from Disparate Information Systems, *Proceedings of the 1989 NSF Workshop on Heterogeneous Databases*, December 1989, Evanston, IL.

[12] M. Siegel and S. Madnick, "Schema Integration Using Metadata", *Proceedings of the 1989 NSF Workshop on Heterogeneous Databases*, December 1989, Evanston, IL.

[13] M. Siegel and S. Madnick, "A Metadata Approach to Resolving Semantic Conflicts", *Proceedings of the VLDB Conference* (Barcelona, Spain), September 1991.

[14] M. Siegel and S. Madnick, "Context Interchange: Sharing the Meaning of Data", *SIGMOD Record*, December 1991, pp. 77-79.

[15] D. Strong, Y. Lee, and R. Wang, "Beyond Accuracy: How Organizations are Redefining Data Quality", TDQM Report TDQM-94-07, MIT Sloan School of Management, Cambridge, MA, September 1994.

[16] M. Van Alstyne, E. Brynjolfsson and S. Madnick, "Why Not One Big Database? Principles for Data Ownership", to appear in *Decision Support Systems.*

[17] R. Wang, V. Storey, and C. Firth, "A Framework for Analysis of Data Quality Research", to appear in *IEEE Transactions on Knowledge and Data Engineering.*

[18] R. Wang and S. Madnick, "A Polygen Model for Heterogeneous Database Systems: The Source Tagging Perspective", *Proceedings of the VLDB Conference* (Brisbane, Australia), August 1990, pp. 519-538.