Applied Operational Management Techniques for Sabermetrics

An Interactive Qualifying Project Report

submitted to the faculty

of the

WORCESTER POLYTECHNIC INSTITUTE

in partial fulfillment of the requirements for the

Degree of Bachelor of Science

by

Rory Fuller

Kevin Munn

Ethan Thompson

May 28, 2005

Brigitte Servatius, Advisor

Abstract

In the growing field of sabermetrics, storage and manipulation of large amounts of statistical data has become a concern. Hence, construction of a cheap and flexible database system would be a boon to the field. This paper aims to briefly introduce sabermetrics, show why it exists, and detail the reasoning behind and creation of such a database.

Acknowledgements

We acknowledge first and foremost the great amount of work and inspiration put forth to this project by Pat Malloy. Working alongside us on an attached ISP, Pat's effort and organization were critical to the success of this project.

We also recognize the source of our data, Project Scoresheet from retrosheet.org. The information used here was obtained free of charge from and is copyrighted by Retrosheet. Interested parties may contact Retrosheet at 20 Sunset Rd., Newark, DE 19711.

We must not forget our advisor, Professor Brigitte Servatius. Several of the ideas and sources employed in this paper came at her suggestion and proved quite valuable to its eventual outcome.

Table of Contents

Title Page
Abstract i
Acknowledgements ii
Table of Contents iii
1. Introduction 1
2. Sabermetrics, Baseball, and Society 3
    2.1 Overview of Baseball 3
    2.2 Forerunners 4
    2.3 What is Sabermetrics? 6
        2.3.1 Why Use Sabermetrics? 8
        2.3.2 Some Further Financial and Temporal Implications of Baseball 9
3. Building the Database 13
    3.1 The Necessity of Play by Play Data for Accurate Analysis 14
    3.2 Current Data Storage Techniques and Providers 15
    3.3 The Database Structure/Design 21
    3.4 The C Code 23
    3.5 The New Possibilities Opened by the Database 24
    3.6 Brief Descriptions of the Proof-of-Concept Parsers 26
4. Conclusion and Discussion 28
    4.1 The Current Status of the Field of Sports Statistics 29
    4.2 Examples of Sabermetric Model Development 30
    4.3 Outside Promotion of the Project 34
        4.3.1 A Report on the 3/26 Meeting of the Boston Chapter of SABR 35
        4.3.2 Another View of the 3/26 SABR Meeting – Ethan Thompson 39
        4.3.3 The 4/16 Meeting of the Connecticut Chapter of SABR 40
    4.4 Potential Future Work 42
Appendix A: Retrosheet File Parser Code 44
Appendix B: Retrosheet Event File Sample 56
Appendix C: Handout at 4/16 SABR Meeting 61
References 63

1. INTRODUCTION

The famed Yankees catcher Yogi Berra once outlined the complexity of baseball, "Baseball is ninety percent mental. The other half is physical." He couldn't have been more right. In this era of highly tuned players and escalating salaries teams and players are trying to find an edge over the competition. That edge can come from off the field preparation, studying the game and understanding what the capabilities of a player truly translate to for a team. Baseball's physical aspects readily present themselves for analysis, but often the data is unmanageable.

The purpose of this project is to make the data manageable. We will create a user friendly data management system that will facilitate future analytical approaches to baseball. Recently, many statistical models have been created to evaluate players and determine if old conventions were accurate. These models are most readily known as sabermetrics because the effort to create them was spearheaded by members of the Society for American Baseball Research or SABR. The most popular example is the OPS (On-base percentage plus slugging percentage), which is a more accurate measure of a players contribution to a team's offense than strictly on-base percentage or slugging alone because it accounts for the impact of each time a player gets on base. The proper usage of such statistics can allow teams to gain a great advantage, but researchers attempting to employ them must first have access to a way of working with and storing the data they need.

Commercial systems exist; however, it is cost prohibitive for researchers to invest in such programs. We hope to provide a no cost alternative that will further the analytical approach to baseball. Over the past thirty years the field of sabermetrics has risen from a handful of amateurs self-publishing books and advertising them in magazines to a healthy, wide-spread network of researchers. However, the data needed to perform the mathematical analysis required to find new knowledge can be difficult to work with, being found in obscure file formats or being held by companies which charge large sums of money. Hence, the database exists to allow that wide base to properly examine the data in baseball.

2. SABERMETRICS, BASEBALL, AND SOCIETY

2.1 Overview of Baseball

The IQP serves as a bridge between society and the skills students learn in school, asking students to use their skills to create or examine some aspect of that society. Therefore, a project's subject must be carefully chosen so as to examine something which is of demonstrable value to the community. This project intends to aid in the analysis and quantification of the sport of baseball. Baseball may at first seem a frivolous subject for a project, but in truth baseball occupies a central pillar in American life, culture, and economics. Herein will be a small brief on baseball's effects on and reflections of culture, concentrating primarily on evidence from essays in *Cooperstown Symposium on Baseball and American Culture*, edited by Alvin Hall. [1]

Marit Vamarasi writes of baseball as being "a metaphor of life." He remarks early in his paper on the purpose of metaphors, that being to clarify some misty concept by the usage of ideas more familiar to an audience. Vamarasi goes on to illustrate how the concepts of baseball have become pervasive metaphors in our speech. He gives a series of direct and obvious examples; "made the right call," "bat an idea around," and "ballpark figure" are but a few. He also gives a few less direct but more enticing metaphors, particularly "hit-and-run." The phrase refers in baseball to setting the runners in motion even before the batter scores a hit, but has also come to refer to hitting another car and driving away immediately. Vamarasi speaks of how hit-and-run is a baseball phrase so old and thoroughly engrained in our culture that it has practically lost its original meaning. His most telling point, from this point of view, is the word "hit." Vamarasi devotes little attention to it on its own, but consider for a moment: hit movie, hit single, hit TV show... These phrases are fundamental aspects of the language, and derive directly from baseball's popularity with America.

Such a popularity can be found in American writings in the early 1900s. Harold Seymour's essay "Baseball: Badge of Americanism" spends four pages listing incidents from the 1900s to the 1930s where prominent figures, including Franklin D. Roosevelt, Massachusetts Governer James Curley, and political machine figures such as the notorious Boss Tweed, speak out on the relevance of baseball. He relates how the New York *Globe* once printed articles on baseball ahead of reports on the presidency, and just before that Seymour mentions an anecdote regarding American citizens' preference for World Series knowledge over the 1908 election results. The sport even managed to attain privileged legal status. By the 1930s, he notes, "the courts had established that baseball park owners weren't liable for a spectator's injury if the fan voluntarily chose to sit where there was no screen…"

Seymour also reviews the primary virtues seen in baseball by early twentieth century Americans. He lists several, including the aesthetic pleasure of watching skilled athletes and nostalgia for the fans' own childhood games in the sandlots and parks of the nation. Later on, he provides his thoughts on how baseball changes with American society, listing a number of affluent and flashy aspects of modern play, such as complex scoreboards, artificial turf, lavish skyboxes, and high admission cost. He contends that these reflect modern American society, which after World War II became increasingly rich and focused on "amusing ourselves to death." His point is quite valid. As American standards of living have risen, skyboxes and good tickets act as status symbols, while the flashy scoreboards and turf make the game more colorful and elaborate. Baseball can act therefore as a kind of cultural barometer, offering an incisive look into the attitudes of its society.

2.2 Forerunners

No project can occur without a basis for work. The work herein is no exception. Many men spent long years of their lives investigating the mathematics behind baseball, and many others have recently begun to put that science into practice. A look at some of these men appears to be in order, as it is their shoulders on which this paper stands.

We owe a great debt of inspiration to the book *Moneyball*. Penned by Michael Lewis, *Moneyball* told the story of the Oakland A's in 2002, covering their draft and a fair portion of the season. General Manager Billy Beane featured prominently as he attempted to reconstruct the way baseball teams are run in the face of great opposition from the press, the public, and most of all, his own scouts. Beane employed the work of young men such as Paul DePodesta, whose analysis of baseball statistics allowed him to select skilled young men with consistently high performance when all existing baseball tradition implored him to leap for flashy, brilliant high school boys without a proven track record. As one of the latter himself, Beane understood intimately the ways these boys can fail.[2]

DePodesta featured as a kind of supporting character in *Moneyball*, but at the end of that story, his was just beginning. Having graduated from Harvard and never really played baseball, he used his acumen with numbers to the great benefit of the A's. After Moneyball ended, DePodesta moved on to a new position: he is now the General Manager of the Los Angeles Dodgers. While he shared Billy Beane's hardcore anti-tradition stance, especially the distaste for high school drafting, DePodesta currently allows Louis White – a member of the opposing school of thought – to run his scouting operation. DePodesta notes in a February 24, 2004 interview with Jim Callis for baseballamerica.com that "different circumstances call for different solutions." The interview mentions that, with so much more cash available to him than with the A's, DePodesta may be more willing to take long shots on the off-chance they deliver.

Another currently practicing advocate of sabermetric management is Theo Epstein. Epstein came out of Yale, bringing with him a varied background including psychology, philosophy, political science, and American studies (his degree actually being in the latter). He proceeded to the University of Chicago, where he attained a law degree. He was only thirty last year when the Red Sox achieved victory in the World Series for the first time in 86 years. Yet, novelist Leslie Epstein noted, at that age, "Alexander the Great was already General Manager of the world." Appropriate, as Epstein was responsible for acquiring players such as David Ortiz and Curt Schilling, whose performances contributed greatly to the 2004 victory.

Epstein revealed his way of thinking in a session with Dennis Tuttle for The Sporting News. When asked his thoughts on statistics, Epstein stated he thought batting average was one of the most overrated statistics available to the sport, as it didn't actually measure value to the team very well. He then remarks that he would like an accurate defensive model, something elusive to this date. This is something this project's database can be used to develop.

Maybe most important of all the past sabermetricians is a man by the name of Bill James. James writes one of the most authoritative reviews of baseball statistics, *The Bill James Handbook*. The book reveals vast amounts of statistical data and formulas while simultaneously giving new insight into the manner in which the numbers are analyzed. James started his work with the *Bill James Baseball Abstract* in 1977, publishing from home, selling 75 copies. The next year, he sold 250. In 1979, he sold still more, and began attempting to create his own new statistics, in the form of the "Runs Created" formula. The formula was not great, but it got many others thinking, some of whom are mentioned in *Moneyball*. It has also been optimized over the past few years, resulting in a definite improvement in its accuracy. James had begun a process by which many new statistics would be created, including the On-Base Plus Slugging measure and Win Shares, which James wrote an entire book about in 2002.

James ended his run of the *Abstract* in 1988, but continued publishing hardcover books and his Handbooks. He worked for STATS, Inc., providing data and analysis to media organizations until 2001, when Fox bought the company. The Red Sox hired him the next year as a special advisor, and it is rumored that he was involved in some of Theo Epstein's decisions, such as the Ortiz acquisition.

2.3 What is Sabermetrics?

Sabermetrics has walked hand in hand with baseball since the sport first gained fans. Taking dozens of forms, from fireside chats to newspaper articles to scholarly papers, it represents the work of thousands over the course of a century. Yet, what exactly can be called 'sabermetrics?'

The best known definition of sabermtrics comes from Bill James, whose Baseball Abstracts and other writings have had a massive impact on the field. He defined it as “the search for objective knowledge about baseball.” This definition is technically accurate and does encompass the field, but it does not describe the field in any useful detail. No methods can be found in it, nor can a reader determine exactly what kinds of people perform this work. Thus a more thorough definition must be constructed to aid those unfamiliar with the field in understanding sabermetrics.

Sabermetrics focuses first and foremost on the statistics of baseball and what they mean. Modern baseball records literally hundreds of thousands of pieces of data every year, and assembles them into the traditional statistical line-up that can be found on the back of any baseball card. The sabermetricians look at these statistics and ask the question, “What does any of this information actually mean?” The goal here is to look at, say, batting averages and determine what a given batting average actually means for a baseball team, and then use that knowledge to improve the ability to predict a player’s contribution to a team. Such a goal by necessity implies a certain degree of statistical analysis. Sabermetrics can be considered a subfield of statistics focusing on problems in baseball, much like other areas of statistics concentrate on industrial efficiency or the like.

Of course, this can lead to unexpected conclusions. Perhaps the sabermetrician finds that an existing statistic doesn’t actually serve any purpose, or notices an aspect of the game that does not currently exist in a statistically described form. At this point sabermetrics assumes its second task, that of building new statistical models to quantify some aspect of baseball. These new models may or may not grow into something of true import, but they represent advances in understanding the game all the same.

This latter point requires a bit more explanation. Sabermetrics, in addition to simply including statistical analysis, also represents a philosophical shift away from using ‘experience’ as a judge of what a player is able to do. Baseball scouts, relying on their first impressions and their years as scouts, find certain things to be important, things which have recently been coming into conflict with what statistical analyses show to be true. The math-based methodology behind sabermetrics is hard to question, even if individual statistics may prove faulty, and the field grows in size and importance every year, but the end effect on baseball is yet to be seen.

2.3.1 Why Use Sabermetrics?

Why is it necessary to dig deeply into the pile of numbers surrounding baseball? What possible reason could anyone have for throwing years of their life and potentially a fair deal of money at the problem of generating an accurate predictive model? To answer this for the thousands of individual members of SABR would require several years of my own life, but there remains another group whose interest is rather more easily quantified: baseball teams.

A baseball team operates, like any other organization, on a certain income every year. It cannot exceed this income without going into debt, and as this income is what it uses to purchase and retain players, this income is the team's hard limit. Thus, like with any purchase, a team wants to maximize the amount of return for any purchase. As a person refuses to buy a fifty thousand dollar bicycle, so a baseball team refuses a player who demands huge sums but trips on his shoe laces in the field.

This fact has always been true. However, there is a second factor now intruding into the game: rising player fees. Michael Lewis notes in his book *Moneyball* that over the past few decades, the average salary for a player has risen by a factor of three or four, with the biggest stars being paid sums in the tens of millions. This is not a healthy state for a baseball team. The cost of a player is now such that each one represents an outlay of millions of dollars. According to the MLB Player's Association, as of 2004, the average for a Major League Baseball Player is $2,372,189 per year. [3]

For an extreme example of this phenomenon of rocketing expense, the Texas Rangers once signed a contract with Alex Rodriguez for 252 million dollars over ten years. As an interesting comparison, the F-15C Eagle fighter flown by the US Air Force has a unit cost of about 43 million dollars. For the cost of this one man's contract, the American government could purchase five advanced supersonic warplanes and still have enough cash left over to buy half a dozen Abrams battle tanks. If the Texas Rangers had spent that money a different way, they could have bought enough military firepower to defeat a small African nation. Instead, they purchased the services of one man.

For further emphasis, instead apply the logic of cost to infrastructure or construction. The Sears Tower, constructed in 1973, cost $150 million dollars. [4] Baseball is not a game of chump change.

What does this mean for a team? With costs like these becoming increasingly common, managers must be ever surer that the million-dollar man they are acquiring is worth the cost. Statistical measures, which allow the prediction of a man's performance years down the line, are the answer. Since a team's ability to win is dependent mostly on the ability of a player to score runs, and run-scoring capability can be computed statistically, proper analysis allows a baseball team to look at its prospective draftees and combine both their costs and their win-scoring abilities to create a team which will win as many games as possible.

This new ability also allows managers to look at their existing players and determine which have been overvalued. In turn, they can use their men to trade for players who have been undervalued by opposing teams. Lewis discusses in *Moneyball* how Billy Beane uses this tactic to improve the Oakland A's. It allows the manager to improve his own team's status in terms of resources gained per unit of money put out while decreasing the other team's ability to fight back in a game. It is, effectively, sabotaging the enemy with his own misconceptions about the game, and is just another reason why statistical analysis can be of great help to a team.

2.3.2 Some Further Financial and Temporal Implications of Baseball

Overall baseball attendance in 2004, according to ESPN.com, totaled 73,010,839 people. This is most likely a low number, as it would not count guests who got in without tickets, such as children below the age of three. In any event, the count should be reasonably accurate for the purpose of this paper, both because young children are not a major component of the audience and because the counts below are interested in the money brought in by baseball, which infant guests will not increase. For purposes of comparison, there are about 193 nations in the world. Only fifteen of these have a population greater than 73,000,000.

Simple attendance of parks does not represent the only personal-scale investment of time in baseball, as the majority of the sport's fans enjoy it via television or radio. ITMS Sports notes that 52,000,000 people watched baseball on television just in April of 2002 alone. Stretch that average over six months, and you have over 300 million people watching the sport. [5]

A Major League Baseball game requires roughly three hours to finish, on average and barring abnormalities in the cycle of play. The numbers above, summed up, represent the rough equivalent of 380,000,000 people spending three hours each watching the game. This sums to nearly 1.2 billion man-hours. Using a forty-hour work week, those people have expended time equal to 30,000,000 people working for a full week on a project, or the output of the entirety of the nation of Suriname for a full year.

More than that, all of this requires money. Sports Illustrated reports that the average cost of a baseball ticket in 2004 was $21.17. [6] Using the 2004 figures for attendance, a total economic investment of $ 1,545,639,461.63 was made by fans entering baseball parks. Note that more popular teams such as the Red Sox and Yankess boast both higher attendances and higher prices, signifying this estimate may be low. This number is of a similar magnitude to the GDP of small nations such as Suriname or Andorra. [7]

Now move along from ticket sales to the sport's other sources of revenue. In the process of the project the authors attended a meeting of SABR held at New Britain Stadium in Connecticut. Personal observation revealed that a single hot dog cost $2.50, drinks were on the same cost level, and other food items were more expensive: a piece of fried dough for $3.75, hamburgers for the same, and so forth. As a piece of evidence of what this means for the stadium's revenue, every single person attending the SABR meeting appeared to have bought at least one item of food or drink, and in several cases an individual bought several items. Purchasing a hot dog and a soda costs five dollars. Alcoholic beverages such as beer were also served for high prices. All of these figures, it must be noted, are for a minor league game, not for a Major League game.

If we examine the attendance figures above, we find that should only half of the attendees buy a single hot dog and no drink, over ninety million dollars flow into MLB coffers, and given direct experience this figure can be considered a gross underestimate of the actual sum.

After this comes licensing fees from merchandise. The ubiquitous baseball caps with team logos ought to generate an idea of the level of money going in that direction, but to make the numbers more concrete, a short tour of MLB.com's online store revealed prices for baseball caps averaging between fifteen and thirty dollars. Any Major League game audience can be seen to have thousands of attendees wearing such hats, to say nothing of the giant foam hands, team cups, team shirts, shoes, and assorted other paraphernalia. Add millions more in revenue to the mounting sum.

All told, the sport is looking at many billions of dollars of income in a year. By comparison, movieweb.com's list of top domestic grossing movies of all times shows that the number one item, *Titanic*, has grossed a mere $600 million. [8] Major League Baseball is equivalent, each and every year, to multiple Titanics being released by the same company. But this is not the entire story. Baseball is major sport in other nations as well. Japan, for example, has two professional leagues with six teams apiece, and a number of Japanese players have become stars in the U. S. as well as at home; Hideki Matsui and Ichiro Suzuki are excellent examples of this. [9]

This data finally leads to a linked pair of conclusions. First, even completely ignoring its position in the culture, baseball cannot be disregarded as a frivolity because of its enormous economic effect. If baseball were limited to sand lot games by junior high students, it could perhaps be deemed unworthy of study in a mathematical sense. However, baseball actually entails billions of dollars of revenue and millions of man-hours by working-age citizens, expenditures whose analysis could greatly aid economists. While this particular paper is not about the money flow directly, it does factor into the second conclusion: if baseball is so important and takes in so much money, does it not follow that its owners should be interested in increasing the quality and efficiency of play? Would not a true picture of the capabilities of the players allow the owners to properly value a player, thus allowing him to be purchased for the right price and placed where his capabilities truly lie? It is to help lay the foundation for such analyses that this project exists.

3. BUILDING THE DATABASE

This project sought to create an electronic database which could then be used by fans and researchers to perform novel research. Current data storage, as implied above, suffers greatly from a number of flaws, primarily in the realms of accessibility and ease of manipulation. To correct these deficiencies, we chose to create a data system using the file format of a commercial software package, Microsoft Access, in the hopes that this system would be easier and cheaper to employ than existing means of working with the data.

Several fundamental hurdles needed to be surmounted before the actual design process could begin. The most fundamental was the choice of medium, that is, whether the data would be stored on paper or in digital form. While the very goal of the project inherently meant that the database would be digitized, it is important to examine the reasons behind this choice, as they certainly apply to other projects involving large data sets. Indeed, this is a primary reason behind the existence of large archives of research journals on the Internet, widely known examples being Pubmed, JSTOR, and ScienceDirect, all of these accessible via the Gordon Library Database Index. [10] Access was chosen as the program base rather than writing an entirely original database source because Access, as part of the Microsoft Office software package, can be found as a part of most Windows systems. Worcester Polytechnic Institute, for example, provides Microsoft Access on almost all public systems and students and faculty can utilize the software license agreement program to cheaply acquire their own copy.

The next choice revolved around the source of the data in the database. Initially, the choice of sources again revolved around paper or electronic, and again, the decision was made – also as an effective matter of course with the initiation of the project – to use a digital source. However, numerous sources of electronic data exist already, all unsatisfactory to some degree and most of them completely unacceptable. In addition, the team needed to find a source of play by play data; section 3.1 holds details on the reasoning for this choice. Section 3.2 details the process of deciding on a source.

With the data in hand, the final step was to design the database itself. An overview of the design process itself can also be located below, in section 3.3, and parser code involved can be located in Appendix A.

3.1 The Necessity of Play by Play Data for Accurate Analysis

Currently, data exists to analyze baseball statistics much more thoroughly then they are currently studied. Few defensive models exist beyond the basic stats of the outfielder, and catcher models are just starting to be developed. The most detailed analysis has been done on batting statistics, though there is a significant amount of disagreement over which models should be used. Some effective models exist, though they are not yet widely used and do not use all of the data that is collected. Play by play data has the potential to greatly enhance the accuracy of all baseball statistics, though is in an inconvenient format to do so.

The play by play data that must be used to model more detailed statistical models is currently stored in relatively unusable text documents. These files contain data for years of games, but they must be manually searched for any relevant information, making quick calculations almost impossible. The existing models were developed off of less detailed accounts of player performance that primarily included the number of hits, bases, outs, etc. Play by play data includes information on where the ball was hit, who recovered it, and where the ball was thrown after it was recovered. The text nature of this data, however, makes it practically unusable for statistical purposes because there is no way for the information to be analyzed or easily manipulated.

Play by play data is predominantly used now for analysis of specific games or players, and used to compare only a handful of excellent players. While this is an excellent use from the historical standpoint, it is of little use for statistical analysis, where the goal is to develop a model that can be used to compare players not only to each other, but to the league, and develop a model for predicting who will win games and why. Since the goal is to analyze entire years of teams, viewing the data as a series of individual events is very time consuming and prevents statistical analysis.

In order to be used for future statistical analysis, the data must be converted into a format which can be easily manipulated, searched, and graphed. If mathematical operations cannot be preformed on the information, it is close to useless for statistical modeling. If the data requires too much time and effort to search through, only the major players whom people are interested in will be studied, which does not accelerate future studies. If the information cannot be viewed, researchers will have no way of telling if they are developing valid models. Therefore, the solution is to incorporate play by play data into a database such as ours.

3.2 Current Data Storage Techniques and Providers

Data for baseball is available from two primary sources, companies which collect it in order to sell it back to media, baseball and research organizations, and private citizens who collect the data and display it freely, generally employing the Internet as a distribution mechanism. In addition, the data can come in two general formats, electronic or paper, with varying sub-types of data that each present their own virtues and flaws.

Before examining the distributors, the method of distribution must be analyzed in order to provide a proper perspective on the choices and contents made. Baseball data most commonly comes in the form of books. The most famous current book is the *Bill James Handbook*, which descends from the *Bill James Abstracts* discussed earlier. A spiral-bound edition, the *Handbook* contains hundreds of pages of tables and mathematical data. The online group baseballprospectus.com, which publishes a daily newsletter, also publishes a yearly statistical guide for fans and members of fantasy baseball leagues. [11] Simple searches of online book stores such as Amazon.com revealed other books, such as the *Complete Baseball Record Book* [12] and the *Elias Book of Baseball Records*. [13]

No dearth of written statistical guides exists, yet paper publication of statistics possesses several key flaws. Intrinsically, the data becomes more difficult to manipulate. Anyone wishing to perform an analysis of large amounts of numbers, such as a hitter's performance against every single left-hand pitcher across five years of play, must either transcribe the numbers into a computer program or resign himself to working out excessively long math problems with enormous data sets. Over the course of a baseball game a batter must, at minimum, face a pitcher three times, for there are nine players in a hitting rotation and twenty-seven outs in a game. A baseball season has over one hundred and fifty games, so even if a player is only present for half of a season due to injuries or rotations with other players to keep each healthy, the data set for his at-bats contains 225 entries. This assumes that the team *never* gets a hit, a ridiculously unlikely occurrence. Only 256 no-hitters have been recorded in professional baseball since 1876. [14] This averages about two per year. One hundred and fifty no-hitters in a year against one team defies probability, so the dataset for any hitter will presumably include far more than two hundred entries, and the average proves to be more like 400-600 per year. [15]

As another strike against them, hardcopy publications cannot be updated with the same ease as an electronic format. Books which cover changing situations, such as encyclopedias or baseball statistics depositories, must deal with the fact that its information will be out of date within a year. Encyclopedias inevitably carry the year of their publication, and so too do every one of the books mentioned above. Each is either suffixed by the year of the data it covers or contains an edition marking somewhere prominent on its cover. To add more data to the store, a new book must be compiled and released, which is expensive in addition to inconvenient.

Third, paper storage requires space. The various baseball statistical journals above tend to be thick: the Bill James release for 2005 is 416 pages, and ballprospectus.com's offering nearly reaches 600 pages. To have a mere twenty years of data from the century and a half of baseball would, given a book with an average size of four hundred pages (four hundred chosen instead of five hundred because earlier books, due to shorter seasons and fewer teams, would have fewer pages), mean 8000 pages of text. Searching through such records requires vast amounts of time and energy, and doing any sort of large-scale analysis beyond simple examinations of win rates and batting averages becomes a daunting task indeed.

By contrast, an electronic system presents numerous tantalizing possibilities. By presenting the data to the end-user in an already-digitized form, it allows manipulation of numbers far more rapidly and easily, particularly if the data comes in a format which already has access to calculation packages, such as a data file for a spreadsheet program. The computer can also allow more complicated calculations on a grander scale than would be practical for a researcher to create.

Electronic data can be rapidly updated via the Internet. Using the above example of an encyclopedia, the Microsoft Encarta Reference Library provides the ability to update articles on the fly via an Internet connection. This enables the user to ensure accuracy on his topic of interest. [16] In the case of a baseball database, the data could be updated in a similar fashion, or by downloading a newer, updated copy from the server if the product is publicly distributed. Furthermore, such a database could be updated on the fly. If an organization, such as retrosheet.org (described below), possessed a large group of workers, it could conceivably record and enter into the database the data for a baseball game within days of it happening. Thus, data could be gained within days, instead of waiting for the end of the year. Also, electronic data takes up little space. A single DVD-R can hold gigabytes of data, the equivalent of hundreds of thousands of pages of text. That same DVD-R is a few inches across and millimeters thick. Fifty of them can be stored in a block five inches across and half a foot high, weighing a pound or two at most. Fifty books require an entire shelf and the collection weighs twenty or thirty pounds. The compact nature of electronic storage and ease of transport is another great advantage over paper.

The superiority of an electronic database becomes evidence when compared with paper. With this established, data must be attained. Designing a system to manipulate baseball statistics without having baseball statistics to input into it would be an exercise in futility, akin to designing a suit without bothering to measure the customer. Therefore, an overview of the existing providers, as implied previously, is in order.

One of the largest official organizations is the Elias Sports Bureau, the official statistic collection service for Major League Baseball. Founded in 1913, the organization was hired in 1922 by the President of the National League to maintain a database of stats for the sport, and the American League mimicked its rival several years later. Currently Elias also holds a monopoly on official statistics for the National Hockey League, National Football League, and the National Basketball Association. [17]

Elias Sports Bureau possesses a web site at esb.com. Unfortunately, they do not offer an electronic database of any sort. In fact, the only information at this site is a brief description of the Elias Book of Baseball Records mentioned previously and an order form allowing one to order a copy. Elias does not provide any electronic data at all, or rather, if it does, it does not advertise it in a readily accessible way. Also, as Elias sells its data to major media groups, it was deemed unlikely that putting it in a database and freely distributing it would be allowed by the company. [13] Another company providing statistics is STATS, Inc. This company, once graced by the presence of Bill James, provides rather more on its website than Elias. Unfortunately, every one of its products is also paper-based. [18]

Very few companies exist to provide the information, and it occurred to the project group that an online solution might prove more viable. This yielded far more profitable yields. Numerous databases of baseball data exist online. Each suffers from strengths and weaknesses with regards to sabermetric research.

By far the most common type of online database, a simple list of player stats can be found most anywhere. Many of these are slickly done, with interactive interfaces and vast amounts of data at their disposal; others present no relevant data to a sabermetric researcher. Baseball-reference.com presents a wonderful example of this sort of resource. It provides data on players going back over a hundred years, with full career information. The data is presented in clean tables, and while the site has little fluff, there is little need for frivolous materials or graphics. By contrast, the data at bballsports.com is of almost no use for math. The database contains only information on what teams a player claimed membership every season. There are no numbers in evidence. [19, 20]

It might be thought that with baseball-reference.com in existence, this project would not be relevant. Unfortunately, it actually suffers from much the same problems as a paper book. While its HTML nature makes it easy to search and convenient to access from any location, the site also does not allow any manipulation of its numbers. This effectively precludes mathematical usage unless the data were to be parsed into an existing program, and this would be made inconceivably difficult due to the literal tens of thousands of baseball players over the decades of the sport, for baseball-reference.com represents each player as a separate HTML entry. This site is impractical as a data source, although it succeeds as a reference work; its lack of play-by-play data is regrettable but all too common.

Two other potential sources revealed themselves over the course of the investigation: a database created by Sean Lahman, and information from the organization known as Retrosheet. These two candidates represent the best options available to the group.

The Lahman database exists at baseball1.com, in purchasable CD-ROM or free downloadable format. The default download is a nine megabyte .zip file containing an Access database file, but it can also be found in comma-delimited and Excel spreadsheet versions for those without the Access software package. As of May 2005, Lahman's database contains data on all players from 1871 to 2004. The database includes large amounts of biographical data and also includes a number of sub-tables for batters, pitchers, and composite team accomplishments. As an example of comprehensiveness, the batters table contains at-bats, runs, hits, doubles, triples, home runs, RBIs, stolen bases, times caught stealing, base on balls, intentional walks, strike outs, sacrifice hits and sacrifice flies, times hit by pitches, and times grounded into double plays. This data is broken up by year as well, so that only one year of a player's performance can be examined. [21]

The Lahman database, as it stands, appears to be exactly what this project seeks to create. Unfortunately, it has its own limitations. The most important flaw, from the perspective of a potential researcher, is the lack of break-down of statistics. While the database serves well for calculation involving overall trends in the sport or the majority of a player's career, it cannot be an aid for smaller calculations, such as whether a particular player performs well against certain kinds of pitcher. In fact, it will not even serve to help in certain kinds of large-scale calculations. For example, the database does not record tendencies for batters to hit to certain areas, which could be of great aid in determining the effect of pitch choice, batter control, and fielder position on reducing number of hits. Lahman does not break down numbers by game or play either, giving season wide statistics.

Retrosheet does not provide a simple database such as Lahman's. Instead, Retrosheet, a volunteer organization, provides a vast quantity of data in the form of individual files found at retrosheet.org. [14] Its major resources consist of a large compilation of box scores from 1871 onward and, most importantly for this project, a series of "event files" for a block of seasons roughly twenty years long, starting in the early seventies (uncertainty regarding length is due to several incomplete years on both ends of the spectrum, for those years are still being added to the archive). Event files represent logs of entire baseball games. These logs are not just standard information about hits, strike-outs, and walks, but rather consist of play-by-play descriptors of every single event in the game that can be usefully recorded, including the location where a ball lands (based on a grid system which can also be located on Retrosheet's event files web page), position that fields the ball, position it is thrown to, and every position involved in the play. The data is recorded as numbers, and though you may have the position, the player is not known. Cross-linking with other data on the site, the players' names can be determined.

This data appears initially perfect. After all, if a sabermetrician has access to all plays in the course of a game, he can perform practically any analysis he so desires on individual performance, and the location-of-hit data can certainly help in the creation of a defensive model. However, like all other previous sources of information, Retrosheet has its own flaws. Retrosheet's event files must be downloaded en masse, for there are several hundred of them, each file corresponding to one team's games for one year. Furthermore, the data itself can only be read in two ways, Retrosheet's provided command line interface program or via directly examining the files with a text reader.

The command line program can technically convert an event file into a format suitable for conversion into a database program, but this will result in a text file with the data inside, which is what is already technically present. Furthermore, the program itself will not run on a standard Windows XP machine, and as this system has become the standard machine in the personal, academic, and business world (for example, all general-use computer labs at Worcester Polytechnic Institute employ Windows XP), the program becomes harder to employ. Measures can be taken to activate the program, but these involve utilizing the relatively ancient MS-DOS command line.

Directly reading the text presents another problem. Rather than employing words, a segment of data representing a game played by the 1971 Atlanta Braves has been included as Appendix C; due to the nature of this extract, including it directly in the body would unacceptably break the flow of the paper.

Keeping that example of a Retrosheet file in mind, consider that a team plays over one hundred and fifty games in a season. With about twenty teams in play at any point, the resultant thousands of pages of data would cause massive information overload instantly, and even that ignores the obscure set-up of the data itself. Therefore, the data had to be processed heavily in order to become eventually become a viable database.

3.3 The Database Structure/Design

Between 1974 and 1992, the dataset used from RetroSheet.org, no fewer than 3,400,000 separate plays were recorded. This staggering number meant we faced some difficult design decisions about how to store the data. First and foremost, all the useful data we could extract needed to be accessible in some way or another. Secondly, search times needed to be relatively fast, for all that databases are efficient, we were entering the sizes even databases have trouble searching. Thirdly, we needed to minimize space usage.

The first idea that came to mind is to simply take, for every of the 3.4 million records, all the data associated with it and dump it into one large table. However, this table would have over 50 columns in it, and at close to 10 letters per column on average, this would take 1700 million bytes of hard-drive space, over 1.5 GigaBytes (GB) for a theoretical minimum size! And because the size was so large, while it would be nice and simple to ask any query of the table, it would take a long time to look though the whole thing.

Since some of the data was very repetitive, the data concerning the entire game for example, we decided to split off a table for game data: we called it the Info table. We shortened the GameID, already listed in the raw data, to Gid, and used it as a link between the main table (the Play table, as it contained all the individual plays) and the Info table. With the Gid serving as a key for the Info table, we now had a way to access any of the data we needed about a game, without having to have it around until we needed it. But there was one major trade off, as there almost always is: it took much longer to run. Now, to reference a bit of data buried in the Info table for a query required searching though the entire Info table for the one match for every element in the Play table! But it is rare that anyone will need this data at first level of a query, where you filter out most of the data you aren't working with. And once the majority of the records are discarded because of the first query, then linking into the Info table takes comparable time to manipulating all the data in the first place. (For a greatly restricted set of the data, it would actually be significantly faster, while if little had been removed it would be much slower, with the extreme being where no records had been filtered out of the Play table.)

We did much the same thing with the fielders playing for each position. In a table named Field, we recorded all the different field states, different configurations of the 9 fields (and 1 dh if needed). We removed duplicates and created a key for the table, the Fid, which was added into the Play table instead of putting the actual fielders right into the Play table. There were two Fids needed, one for the home team, one for the visiting team. Just like for the Game table, many of these entries were used many times, and as well as making the queries that didn't reference them faster by breaking apart the data into many tables, the overall size was again greatly reduced. After the database was set up to use these three tables, we found it only took about .5 GB, (slightly compressed) less than

1/3 the theoretical minimum of the other approach. Conveniently, it also would fit on a single CD, making it much easier to distribute physical copies if needed.

What was left in the Play table was the data unique to each play, the data with no logical grouping. The Play table had a two part key, known as a distributed key. The first part was the Gid that linked to the Info table, since it was given in the raw data and was unique to each game played. The second part was implicit in the data (see "The C code") by the ordering of the plays, but databases lack the concept of a "order" of the records contained in them, so the C code generated a Pid (or Play ID) for every play. Within each game the plays occurred in order of ascending Pid, if play order needed to be regained at some point. Year was also re-added to the Play table, because we found we often needed to limit the year we were working on; having to limit year so early meant the database had to link into the Info table very early in the query, causing it to be a very slow query. So though it took more space, we judged the time savings more important in that case.

3.4 The C Code

The data available from RetroSheet.org looks, at first glance, to be ready to be fed straight into a database. But looks can be deceiving, and while it did come nicely comma separated, there were still many obstacles to overcome before the data was in the format the database needed. So a little bit of code was written: a little bit that grew and grew as the demands on it mounted. Lacking a knowledge of Perl, our language of choice was C/C++. (Though, with hindsight, it would have been better to spend the time learning Perl.)

The first order of business for the code was to take the 900 separate files and merge them into a smaller, easily importable number. Though one of each type was our ideal, Access could not handle importing such large files, and we were left producing ten of each type. Irregular file names were also a nuisance to compensate for.

The database format we intended to use required the creation of separate tables for data concerning whole games (info data), data concerning the fielders (field data), and the data for the individual play data (play data). Therefore we split the data off to separate files. The play data required an explicit numeration of the order in which plays occurred in place of the implicit ordering present in the raw data: database inherently have no implicit ordering, all ordering must be explicit.

The info data required more work; there were many irregularities that required filtering out. Some data was judged unnecessary, like the "inputter", some was redundant in a few instances, and some was missing instead of listing N/A or some equivalent. Once we knew which parts we wanted to keep, this consisted mostly of tracking though the raw data to find the irregularities and adjust for them.

Most difficult was the field data. The raw data had substitution events and a sector at the start of the game for the starting players; we needed a list of all the "field states" (state of the defense in the field) that occurred, and a link between each play and the appropriate field state. (See the database design section for an explanation why we used this design) The code tracks, play by play, the field state, updates it (using the batting order, since who is being substituted out isn't explicitly listed) when appropriate, and automatically removes duplicate field states. We have hand verified a few instances to be sure all of that was working correctly, no irregularities were found in any of the cases we looks at.

The final version of the C code is attached in Appendix A.

3.5 The New Possibilities Opened by the Database

Restructuring the data into database format, especially the format described, brings a myriad of benefits. Most notably, all the relevant data about a play is now either directly in the line, or in one of the links from that line. No longer does a program need to simulate the entire game to find out who is playing first, nor is it necessary to go to a separate file for a different team's games. The complete nature of each entry allows the real power of databases to come out: finding elements with certain properties, and counting them. This forms the basis for calculating many varied probabilities, ranging from the simple batting average out too the complex proof of concept query we built involving probabilities for each fielder fielding a ball hit to each of the different sectors.

The other major useful feature of the database is its ability to easily filter out results that don't satisfy one of the necessary conditions for a given problem. These filters can either be to block out events that "don't count" for a given problem, or can be used to subdivide the data to allow modeling of more complex interactions. While always possible, since no data is created, the database makes these types of operations on the data much easier to implement.

Not everything is as we might wish, however: there is a major inconvenience in our current database compared to the ideal system. While we made great progress at interpreting "project scoresheet" data about each play (such as the sector or whether or not a hit was a line drive or a fly), reliably and accurately interpreting all the different variations requires far more time then we could commit to this particular problem. We can get most of the "project scoresheet" data parsed within the database, but even missing a small about can skewer some of the more complicated calculations. Some of the problem is with bad data entry into the raw data files, and there is no fixing these errors short of going back to the original score cards and reentering the correct play. But part of the problem is the complexity of the scoring standard. Hopefully a future group of some sort can take the current "split" approach, which breaks apart the event into a few different parts which then still need to be parsed, and replace it will a fully extended "database readable" version of all the data contained within the event data. We currently can successfully parse out weather or not the play resulted in a hit, but even then we need to ignore bunts and a few other special cases. Every type of question like this one will have its own quirks, many the result of the rules of the game, many more the result of the standard the data was first recorded with. We have provided proof on concept for these parsers, but until a mostly standard set is produced, users of the database will have to reinvent the wheel and implement their own parsers for the less commonly used chunks of the data they want to get at. It is worth pointing out that, for all the trouble this is, this exact same problem was strictly worse for the original data, as all the parsing needed to be manually done, not just chunks of it.

3.6 Brief Descriptions of the Proof-of-Concept Parsers

Here follows a short description of several representative queries in the database and the capability they demonstrate is available to the user. Not all queries are listed here because a number of them are effectively duplications of others, such as the various "number of statistic-x per year" found in the list. The below are simply examples of the overall capabilities of the system. Furthermore, we encourage users to experiment and build their own queries, as the existing set may not be exactly what a user will desire.

Note that the syntax used in Retrosheet files includes a large number of shorthand codes for particular events, and these codes find themselves replicated in the database. Writing an extensive tutorial to help would be redundant, for Retrosheet's own event file documentation can be found at http://www.retrosheet.org/eventfile.htm. Of most interest to the reader is the explanation of what individual event codes mean, allowing users to actually determine what codes to employ in a query.

HomeRuns_Player_Year: Shows the number of homeruns corresponding to a single player ID code per year; the ID chosen here is rosep001, or Pete Rose. The user can substitute any desired ID code or statistic and gain similar data.

Intentional_Walks_Player_Year: Shows the number of intentional walks accrued by all players over the course of the entire period covered by the database. This demonstrates the capability to acquire all examples of a particular statistic in order to perform large-scale work.

Locash_fidhash_play_link: Gathers all play/event info for all games in the database.

Locash_field_id_iif: Gathers field id and positional data.

Locash_link: Merges the above two queries and allows manipulation of all gathered data.

Out select: Displays events where a player was knocked out by a play as opposed to striking out at the plate, walking, or getting a hit. Shows that the program can differentiate between various outcomes and allows users to perform calculations that require this level of discernment. For example, if the user filters the database for a particular pitcher, he can then use this sort of query in combination with others to determine if particular pitchers can control how well fielders can acquire outs.

Outs_to_field_position: Shows how many times per year a given player was thrown out or flew out to a fielder. This is another demonstration that the program can discern between different kinds of field events.

Player_sector_split: Displays the number of hits by each player to each sector of the ball field based on the Retrosheet.org grid system (for a subset of years, in order to show that this ability is also available; the entire database need not be shown).

Sector_hit_Y_n: Displays the number of hits of all types (pop-flies, line drives, and so on) to particular locations on the field for each year of the selected interval. This allows users to examine, for example, whether there are certain zones in the field which attract the ball more often, if these trends change over time, and therefore can be used to properly position defensive players or develop hitting and defensive models which take into account location.

4. CONCLUSION AND DISCUSSION

With the database completed, prudence advises that it be examined to determine if the project has succeeded in creating the tool it desires, and whether the database serves a useful role in the world.

In summary, the database itself consists of an Access file which, when compressed using the.zip file format, occupies 135 megabytes of space on a hard disk or CD-ROM. The file format was selected because both Windows XP and Apple's OS X have the ability to open these files without requiring external software, and furthermore, free third-party utilities such as 7-Zip exist online to allow access to these files under older operating systems such as Windows 98. [22] Extracted, the file requires 836 megabytes. This is suboptimal, as initial plans hoped to allow placement of the entire file on a CD-ROM for easy transport by users or for distribution to interested parties unable to download the file. The compressed form will fit with plenty of excess volume, but unfortunately the act of compressing and uncompressing a file can introduce artifacts and errors. Thus users may experience errors after acquiring the database, requiring a second acquisition or download. Also, though modern hundred gigabyte hard drives can easily absorb the impact of an 836 megabyte file, many users may not have up to date computers. This flaw expressed itself pointedly when it was noted that the laptop of one of our team had only a few gigabytes of space available at any one point, meaning this database would absorb a third or more of his available computing space. Given the original projections of a database of over 1.5 gigabytes in size, twice the current figure, the present contraction does represent a definite improvement. Still, it would be neglectful to not make note of the extent of the file.

The database incorporates all existing Retrosheet event files available when the finalized version of the database was compressed on April 16, 2005. The database's system still relies or functionality on the dense and unintuitive Retrosheet system. On the upside, even if the reader must still skim the Retrosheet tutorial given earlier, the ability to work with all of the data directly and the provided example parser queries should grant users a greatly enhanced ability to work with the data over the complex and murky text files.

As to whether the system can be relevant, review of current sabermetric work and the response by outside parties to the database idea will grant some insight into the perceived usefulness of this finished system.

4.1 The Current Status of the Field of Sports Statistics

Sports statistics continue to gain in relevance in the current day. Beyond the earlier discussion of the work of pioneers such as Bill James, Billy Beane, and Theo Epstein in altering the sports themselves, recent trends in book releases and math education show that sports may also be affecting math.

The magazine Math Horizons serves as a tool for undergraduates to introduce themselves to the world of practical mathematics by laying out applications for math in real life. The April 2005 edition focused on sport-based problems, using a series of problems or situations to illustrate mathematical principles or gain insight into a system. Directly applicable to baseball, Mark Schilling writes the article “Chance Encounters: Sporting Events,” using statistical analysis to look at the significance of the number of seven-game World Series events. The same article also explores whether any actual influence can be exerted on the prevalence of tied-score soccer games. Rebecca Sparks and David Abrahamson choose to explore if, by analyzing the data on a candidate, it becomes plausible to predict how voters will rank potential candidates for an award. They chose to use the Cy Young award, and found that this award, at least, proved to be mathematically predictable. Math Horizons is published by the Mathematical Association of America, and can be found at http://www.maa.org/Mathhorizons/. [23]

Jim Albert takes the idea of educating undergraduates with sports one step further with his textbook *Teaching Statistics Using Baseball*. Albert’s book exists to act as a guide for an introductory statistics course, utilizing baseball as the basis for its case studies and analytical models. Albert reasons that by using baseball, students will have a familiar frame of mind which will help them grasp unintuitive concepts in analysis more rapidly. The course has been taught successfully at Bowling Green State University. Albert possesses a personal site at [24], which contains sample material for a class, solutions to the book's problem set, and the datasets used to create the book's problems.

4.2 Examples of Sabermetric Model Development

The health of sabermetrics lies in more than just the subsidiary uses of the field. Actual mathematical development is necessary for sabermetrics to grow and expand. An examination of the status of model development indicates the health of the enterprise as a whole.

Sabermetrics attempts to develop statistical models for the performance of players. Play by play data has been used create models that can be used to predict how well players will perform. Most of these models have focused on hitting, as run production is seen as the most important aspect of a player by many. This data, however, can be used for more than just hitting statistics, as it includes who was involved in generating outs on the defense. For instance, recently, the data has been used to generate models to evaluate catchers.

Development of more advanced hitting models began in the late 1950s when the early statistics of batting average (hits/at bats), slugging percentage ((singles + 2*doubles + 3*triples + 4*home runs)/at bats), and on base percentage ((hits + walks + hit by pitch)/(at bats + walks + hit by pitch + sacrifice flies) were shown to have little correlation to actually winning games. Most of the models developed give a weight to various plays, and then divide by the number of at bats. Because of the significant research that has gone into hitting statistics, a number of accurate models have emerged. Despite the advances, batting average and slugging percentage are still the primary statistics cited in newspapers, television, and the backs of baseball cards. [25]

Major League Baseball is very slow to grow into new statistics. Very few are actually publicized by the MLB organization, and these statistics are all early, easily computed numbers that are generally considered poor indicators of player performance. The hitting statistics officially supported, batting average, slugging percentage, and on base percentage, are considered some of the worst indicators of players, and are widely believed to lie about how valuable a player is to a team. The pitching statistics reported by MLB are very basic, but are some of the primary statistics for comparing pitchers. The strikeout to walk ratio, hits per nine innings, and earned run average are some of the primary statistics, even though they are relatively basic in their calculations. Little has been done beyond these statistic publicly to compare pitchers analytically. The fielding statistics, defensive efficiency rating, fielding percentage, and range factor, are very poor statistics that can grant no more than a vague idea of how good a player or team is at defense. These calculations are almost pointless for some position, mainly the catcher, and it is widely true that there is no good defensive model.

Going away from the basic models supported by MLB, many sabermetricians have developed models in an attempt to display players in a manner that more accurately portrays ability. The dissatisfaction from old statistics did not begin recently, and in fact began many years before the term sabermetrics was even coined. In the 1954, General Manager Branch Rickey developed a model for analyzing a team's performance. Though not well known, this model was reportedly very effective in calculating the team ranking given all of the data for the year. Unfortunately, it was a poor model for predicting the performance next year. Though it was relatively simple, it could provide very useful information about the weaknesses of a team overall. For instance, teams lacking a hitter that can consistently bat players on base home will have a very low clutch value and would be easily identifiable. The addition of one such player could bring a team with average scores elsewhere from a middle position to being a significant playoff contender. This could be immensely beneficial to managers looking at what the team needs overall, but the statistic would be completely useless in comparing individual players. Research has not been done to determine if the model still holds true and accurately depicts the team rankings.

Baseball statistical analysis hit the mainstream baseball culture in the late 1970s, around the time Bill James and other sabermetricians started releasing baseball abstracts that contained new models for analyzing hitting statistics. A few models were generated in the intervening time, though Bill James's Runs Created model has taken center stage, partly because of constant minor updates to the model in his annual publication of player scores. Many similar models have come into existence, such as the Linear Regression Model presented in *Curve Ball* by Albert and Bennett, or earned run percentage (ERP). In the end, many hitting statistical models have been created, and they produce relatively similar results when ranking players. Though some models may give slight advantages to certain types of hitters, as *Curve Ball* explains that they're Linear Regression model gives slightly different results from runs created when considering fast moving versus hard hitting players, these differences are relatively minor, and the difference between how well the models correlate to actual play is relatively small.

Win Shares is a relatively new method of determining a player's contribution to a team by awarding a player a percentage of the points accumulated by the team. The goal of the statistical model is to create a system where any player's contribution to the team can be effectively modeled and compared, be it fielding, pitching, or hitting, on an even basis. The method was created by Bill James, and is described in his new book title "Win Shares." Some have credited Win Shares as being one of the most effective models out there, and it has been brought up as an effective way for determining MVP and who should be inducted into the Baseball Hall of Fame, because it takes the accurate data for how well the team did. The model is designed in a way that the Win Shares that a team has is directly related to the number of wins of the team. Because of this, it is less effective at predicting how well a team will do in the future, but it makes a good model for showing the results of a season. [26] A team is credited with Win Shares based on the number of runs scored, runs scored against the team, and the number of wins the team has:

(Runs Scored -.52 * League Average Runs Scored) * 3 * Number of Wins = Offensive Win Shares

(1.52 * League Average Runs Allowed – Runs Allowed) * 3 * Number of Wins = Defensive Win Shares

The Win Shares are then divided by player performance among the team. The Runs Created model is modified by the number of times a player is caught out to determine the number of Offensive Win Shares that a player earns.

Defensive Win Shares are significantly more complicated to calculate, and are a function of many factors in which Bill James goes into significant detail to explain in his book “Win Shares.” Factoring in fielding and pitching statistics, the defense is credited with between 25% and 40% of the defensive Win Shares, while the pitcher is credited with 60% to 75%. This value typically is around 30% and 70%. Pitchers are then given Win Shares based on how well they performed (runs allowed versus innings pitched, modified for score in the innings pitched). Different statistics are used for each fielding position, and the positions are ranked based on importance to the team. Bill James essentially selected four characteristics for each position which were important in determining player performance, calculated as a percentage of the team’s total of that statistic. The numbers are then modified by certain criteria, including handedness of pitcher and double or triple play opportunities, to calculate the percentage of the teams Win Shares that should go to a specific player. [27]

In the recent past, a few defensive models have begun to be more thoroughly investigated. David Pinto currently runs a website, baseballmusings.com, where he has developed new models for defensive statistics. Mr. Pinto is a former baseball researcher for ESPN sports center and currently publishes his probabilistic model for range for the last season. The model calculates the probability that an out was created based on the direction, strength, and type of the hit. He splits the field up into vectors that the ball could have been hit into and the area whose responsibility it should have been. It is adjusted for the park, and the handedness of the pitcher and batter. The model only looks at the number of outs created from catching the ball, so does not look at plays that generated an out from catching runners. It is still in its developmental stages, and it conflicts with more accepted models. Mr. Pinto admits that win shares is a more effective model, because it factors in other aspects of defense, though with further development the probabilistic model for range could be very useful in the future. [28]

Two other defensive models, Ultimate Zone Rating (UZR) by Mitchell Lichtman and Defensive Regression Analysis (DRA) by Michael Humphreys, have recently been developed, though neither of the formula is available to the public. The results are posted, though both men are employed by baseball teams, and the formula are proprietary information. Interestingly, Michael Humphreys was not employed by a team when publishing his findings on the DRA formula, though he refused to share the formula in the hopes that a team would approach him for employment, and instead posted his results and compared them to UZR. UZR is based off of play by play data while DRA is a linear regression of more publicly available data and can be calculated for all of MLB history. UZR values for players are no longer computed publicly, because Mitchell Lichtman has recently been hired by the St. Louis Cardinals. [29]

Most of the models described above have arisen within the past fifteen years. The field of sabermetrics clearly retains health, and therefore the database will have some relevance.

4.3 Outside Promotion of the Project

In order to gain a direct indicator of the opinion of individuals in the field regarding the feasibility and desirability of this project, we attended two meetings of SABR, the first to observe and interact with the group and the second to actually perform a short presentation regarding the project.

4.3.1 A Report on the 3/26 Meeting of the Boston Chapter of SABR

Plans had been made by the group to present the database, and the results of the statistical analysis performed with it, at a meeting of the organization SABR, the Society of American Baseball Research. In order to prepare for this, we planned an initial trip to a meeting occurring earlier in the year, hosted by the Boston chapter of SABR. We felt that knowing the general tone and context of the meeting, its attendants, and its presentations would allow us to properly design the talk to meet the standards of the group.

The meeting occurred in Cambridge, Massachusetts, on March 26, 2005. Present were group members Rory Fuller (myself), Ethan Thompson, and Kevin Munn. Pat Malloy, working with the IQP group on an intertwined ISP, proved unable to attend due to medical problems. The group arrived for the meeting a few minutes before its start, introduced themselves to the chairman of the chapter, Seamus Kearney, and took seats to observe the proceedings.

SABR chose to hold its meetings in what seemed to be an old Quaker meeting house near the historical Longfellow building. The chamber selected held room for about forty people. A little more than thirty people showed up for the meeting, including the three IQP writers; the count at our arrival was about thirty, with several late-comers. This left the room feeling comfortably open without seeming empty. It directly abutted a kitchen with a serving counter, and various baseball paraphernalia decorated the room.

Of these individuals, about eight or nine - including our group - were new to the chapter, although one was a transfer from another region. This indicated roughly twenty to twenty-five regular members, allowing for one or two individuals not appearing at that particular meeting.

The attendees were a reasonably diverse bunch, in keeping with the theme that baseball represents a universally American icon. Most of the attendees looked or were older, with Mr. Kearney appearing in his sixties or seventies. The majority looked in their thirties or forties. Our group, of course, rests in the early twenties. Another college-aged fellow attended the meeting, him, like us, for the first time, and some of the regulars appeared to be in their late twenties. Dress appeared to be casual, as no one commented on the jeans and T-shirts we wore. Some of the presenters appeared in good shirts and khakis. Hats, of course, were almost inevitably baseball caps. Notably, a number of women attended. Two of the six speakers we heard were female, and a number of others dotted the audience, including one elderly woman who played in a women's league in her youth.

The mix of people presents an interesting look at the nature of baseball as a sport. The ages of the speakers roughly correlates to what one might expect if we made predictions on attendance of a scholarly meeting on a subject loved by all. No children or teenagers, as youth typically has little interest in research and is more concerned with actually playing the game. Otherwise, we see a rough bell curve, with several young fellows just getting into the field, the larger bulk of the middle-aged, experienced researchers, and the trailing off as age increases and people retire or die. The number and indeed prominence of women at the meeting elicited some thought as well. Baseball, in terms of popular culture, still seems to pass itself off as a male undertaking, but clearly it contains elements which appeal to all, else why would a third of the presenters at this meeting be female? That there is a bias toward males in the fandom is evident, yet simple observation disproves any notions of baseball being restricted to any one group.

The meeting opened with a brief introduction by the chairman, and a few words were spoken by David Southwick about the chapter's 1975 Red Sox Project. The organization intended to catalogue and commemorate the lives and deeds of that particular team, and Southwick reported excellent progress along those lines. As of that date, completion was scheduled for November.

After the introductory speeches, the first presentation began. Joanne Hulbert explored the history and origins of Opening Day in New England baseball as they related to the old holiday Fast Day. As reported by her, Fast Day arose in the initial months of the Plymouth Colony effectively as a response to crippling lack of food. The Puritan leaders of the Massachusetts Bay Colony later implemented it as a means of reinforcing doctrine and helping further the cause of purification of the soul. This worked for some years, but by the time of the eighteenth and nineteenth centuries, most inhabitants of the Boston area considered the date to be a time for field sports and revelry. While church leaders made an effort to make celebration of the holiday into law, the timing of the day - the beginning of April, right as spring arrived - thwarted all such attempts. Citizens continued to spend most of their Fast Day enjoying themselves, and many prominent citizens began to assault the original purpose of the holiday in the papers, referring to it as "Farce Day." It was only natural that in an atmosphere of hostility toward actual fasting and benevolence toward field sports, the pastime of baseball should begin its earliest seasons on Fast Day.

Numerous attempts were made to dislodge this new arrival, including outright changing the date of the holiday, but all were failures, and to this day baseball continues to open on roughly the same date. Similarly, other field sports continue to occur around the same date, such as the Boston Marathon.

Afterwards, lunch arrived and the meeting broke up into conversation amongst the various attendees. We spent our lunch chatting with some of the locals, generally introducing ourselves and explaining the scope of the project to gauge reactions to it. Ethan and Kevin spoke with Paul Wendt. I spoke further with Chairman Seamus Kearney and with member Fred Ivor-Campbell regarding the project. Both men proved quite receptive and friendly, and they willingly gave advice on how to proceed and words of encouragement about the continuing task. Brief words were also exchanged with several others whose names unfortunately went unrecorded, but who nonetheless presented themselves admirably as representatives of the organization.

After lunch three further presentations were held. The first, by Bill Nowlin, was an overview of the research he performed in order to write his book *The Kid*, a review of the early years of Ted Williams' life. Nowlin told a number of stories regarding the trials he underwent finding members of the athlete's family and also went over a number of details on Williams' life. He reported on his conversations with Williams’ Hispanic ancestors and how he had to track down several nephews to actually get useful information about the man’s life. Coincidentally, one of those nephew, also named Ted Williams, did the graphical design for the book.

Next came Cynthia Tan's explanation of the methodology used to write her book about the fifty greatest Yankees' games of all time. The sole Yankees fan at a meeting of Red Sox fans, she caught the expected volume of flak and took it smilingly. Her report listed the criteria she used in an attempt to quantify something traditionally considered unquantifiable. Her intent was to establish a method that would acquire a list of games that represent the true scope of exceptional play, both in victory and defeat. Defeat should be dwelled upon; Tan emphasized that a minimum of ten percent of the games in the book should represent famous losses, as even horrendously devastating failures are remembered twenty or fifty years down the line. She also restricted herself to only one or two games from any particular post-season, and deliberately selected at least half of her games from the regular season. She reasoned that, as the majority of games are regular season, there could not possibly be such a great dearth of memorable games in that category that she would have to devote most of the book to play-off and championship matches. Whether or not she truly succeeded in creating an unbiased list probably cannot be determined, but the method she described was at least an attempt to give exposure to all possible candidates, and for that she gets credit.

Leigh Grossman gave the final presentation we were able to view before being forced to leave to catch the train. He penned a book called the *Red Sox Fan Handbook*. He created this tome to provide an introduction to baseball and the Red Sox for those new to the sport and to try to explain to the newcomer (or fan of another team, perhaps) just why the Red Sox held such a draw for their fans when they consistently did so well right up until the moment their engines failed and they carved a large swath across the ground. His presentation was somewhat shorter than the stated time on the meeting’s program. He gave a few details of the contents of the book, including its focus on the players themselves and not necessarily the history. Grossman explained why some of the questions in the book seemed trivial to hardcore fans, and also noted that the book contained many definitions of things whose meanings fans take for granted and newbies might be unwilling to ask about for fear of appearing silly or insipid. The presentation was basically a large ad for the book, but it was acceptably well-delivered and the book itself seemed an excellent concept for drawing new blood into the sport with less pain than in the past.

After this we spoke briefly with Mr. Kearney, thanked him for his hospitality, and returned to Worcester, Ethan by car and Kevin and I by train.

4.3.2 Another View of the 3/26 SABR Meeting – Ethan Thompson

I arrived first of our group, as I was coming separately and had worried of traffic and parking problems that weren't there. So as I walked in, only a few people had arrived, but over the next half hour, to the slightly late start, more and more people filtered in. Many seemed to know each other fairly well, but when the meeting started Seamus Kearney, the chairman, asked for all new members: asking a few questions of each group to introduce them to the rest of the attendees and welcoming them to the meeting.

Following some chapter business concerning ongoing proejcts, Mr. Kearney introduced the first speaker, and the presentations started. Taking half an hour to a full hour, the presentations were quite relaxed and informal. First covering the origins of opening day, then three book plugs: one concerning the 50 best Yankees games, one about the early life of Ted Williams, and one about The Red Sox Fan Handbook. These were not just plugs for the book, but also detailed descriptions of both what the books contained, and the methodology used in creating them, as well as many related anecdotes. There was a noticeable focus in all the presentations on nontechnical aspects of the game, though none of the presenters were presenting anything relating to the Metrics portion of the organization.

Both during the first break, and then over lunch, discussions of almost anything baseball related seamed to pop up. We were approached by a few different people interested in the ideas we were working on, they all seamed to have advise of some sort (much of it being to talk to someone else in attendance) and all were quite encouraging.

For all that it didn't show in the presentations, these discussions showed a great wealth of knowledge concerning all aspects of the sport, including the statistical analyses. I overheard a quite animated discussion concerning the non-linear nature of skill for pitchers and hitters, how average players of either type can on paper look fairly reliable, but against a strong opponent can be nearly completely ineffective. Even at a baseball focused event, not a typical discussion to be having.

Both the presentations and the social time in-between were quite engaging, and when it came time for us to take our leave (I to family commitments, Kevin and Rory to a tight train schedule) we felt we had found "the group" for the work we were doing. These were the people who drove the search for better understanding of the game, and these were the people that could make great use out of our database. We were glad to have attended, and looked forward to the point where we could present to a similar group our work.

4.3.3 The 4/16 Meeting of the Connecticut Chapter of SABR

We attended a second SABR meeting on April 16, 2005. Going into the conference we anticipated giving a short presentation on the basis of the project and distributing the small data sheet found in Appendix B. The meeting occurred at New Britain Stadium in Connecticut directly before a game between the local Rock Cats and the Harrisburg Senators.

The Connecticut Chapter of SABR had only just formed at the time of this meeting, its second. Most of the attendees were therefore new to the organization. In fact, one man attended because he walked by the group as it gathered outside the stadium and showed interest in the subject. Not all were new to SABR; aside from several transfers from other regions, present at the meeting was Cecilia Tan, from the 3/26 meeting.

Because the chapter is new, the Steering Committee Chairman, Steve Krevisky, has been acting to create a strong relationship between the chapter and local baseball groups, including teams and nearby SABR chapters. As a result, the chapter had secured a private lounge overlooking the field, almost directly behind home plate, and free tickets to that afternoon's game for all attendees.

Connecticut Chapter's mix of attendees resembled the Boston Chapter to a certain extent, with some differences. A smaller percentage of the group was female, and the general mix looked older. Still, about twenty individuals attended; the well-established Boston Chapter had less than twice that, so the new chapter certainly seems vigorous. Several academics were present, including Mr. Krevisky himself, a professor at Middlesex College.

The meeting itself flowed fairly swiftly, with only four presentations. First in the line-up came Cecilia Tan, performing an encore of her Boston presentation. She established her intention to make this presentation in several locations as a kind of publicity tour, and then went into her speech. She did make some changes, altering the trivia questions she asked and passing out numerous small prizes of an unknown sort; our group's vantage point in the cramped room prevented us from actually seeing the prizes themselves, and most recipients immediately stowed them in a bag or pocket.

Unfortunately, the names of presenters two and three went unrecorded. The second presentation detailed the life of one James Henry O'Rourke, a Hall of Fame player from the late 1800s who scored the first hit ever in the National League. The presentation also explained that O'Rourke's house, in Bridgeport, Connecticut, is under threat of destruction and needs to be moved to a different location if it is to survive. The presenter helped organize a small band of locals including a local lawyer, clergyman, and bank executive to raise funds to move and redevelop the house. That group held an annual meeting on April 22 and invited any SABR members to attend, learn more, and decide if the cause of moving the house and making it into a museum on early baseball deserved monetary support.

The third presenter's original presentation apparently became infeasible immediately before the meeting, so he sought to make an impromptu analysis of the nature of a memorable pennant race. It was less of a presentation as it was a group discussion on the merits of various pennant races, with the leader throwing out various facts and using several books as references. Because of its fairly spontaneous nature, the lecture meandered from point to point without actually establishing a concrete base for dialogue and while the presenter brought up several interesting thoughts – he asked, for example, whether the current league divisions in baseball prevented potentially interesting pennant clashes – the talk as a whole cannot be judged to have been a success. With more preparation, however, he can expand his good core idea into something viable.

We stepped up for the fourth and final presentation. Pat Malloy, who worked with us throughout this project as an ISP, acted as primary speaker. The baseball game started very soon after the scheduled time for us to speak, so we were forced to be brief. The speech's basics can be found in the hand-out in Appendix B. We introduced ourselves and our project's purpose, described the database itself, and fielded several question. After this, the meeting adjourned and we settled in to watch the game with the group.

During the first few innings of the game we spoke in small groups with other attendees, including Mr. Krevisky. Again, all we spoke with conferred their interest in our project and wishes for our success. We indicated that hopefully the project would be complete by May. Distributing several hand-outs put our names and e-mail addresses in circulation, and we hope to receive some inquiry regarding the progress of the database at some point over the summer.

4.4 Potential Future Work

The first line of work is to improve the database itself. Additions to the query set and to the data itself will be a necessity to keep it up to date. Fortunately, the first will naturally be performed by researchers as they build their own data models. The second presents more complications. As the data is based on Retrosheet's event files, inputting more data requires more Retrosheet files. Thus, the dataset available may be limited to the existing twenty year span for an indefinite time period.

It is entirely possible that future groups could employ this database as the basis for a project of their own. An example of such a project already exists. Over the last two months of the project (March and April of 2005) Pat Malloy, in addition to his aid on the database project, spent a significant amount of his ISP working on developing a model for accurately analyzing defensive ability. He contacted and employed the aid of Professor Jayson Wilbur in his work. While a complete description of his work is outside the purview of this text it is important to note that such attempts are already underway. Future groups could make use of this database for the verification of existing statistical models or the derivation of new ones.

Other potential projects include rebuilding the database into a new format. A new project could take the existing Access file and convert it into a MySQL-based system with a custom front end specifically for baseball, with built-in tools for calling up common stats in addition to allowing custom queries constructed by the users.

Appendix A: Retrosheet File Parser Code

I, Ethan Thompson, as the primary author of the following code, feel obliged to warn any that might wish to read it that it is, stylistically, a 330 line disaster. We didn't realize the scope of the program and decided to "hack" together the program to get the data ready for the database. In retrospect it would have been better to learn perl and use that, but we didn't know this until we were almost done.

The code:

```
#include "iostream.h"


//syntax:  strsep(char **the_string, const char ',');
//returns the first part, stringp is remainder


char imput[100];
char line[100];// = new char[100];
char path[100] = "/Users/et/Desktop/BASEBALL/1974al/74BAL.eva";
FILE *info;
FILE *field;
FILE *play;
//FILE *sub;
FILE *com;
FILE *data;
char gid[100];
char* swap;
char** thing;
char* temp;
char* work;
int pid;
int inum;
int teamval;


char force[29][20] =
{{"visteam"},{"hometeam"},{"site"},{"date"},{"number"},{"starttime"},
     {"daynight"},{"usedh"},{"umphome"},{"ump1b"},{"ump2b"},{"ump3b"},{"umplf"},{"umprf"},
     {"howscored"},{"pitches"},{"temp"},{"winddir"},{"windspeed"},{"fieldcond"},{"precip"},
{"sky"},{"timeofgame"},{"attendance"},{"wp"},{"lp"},{"save"},{"gwrbi"}};

char fs[2][11][10];
char big[100000][11][10];
int fid0=0;
int fid1=0;
int maxfid=0;
int togval;

void findfid();
void findfid(){

        int check=0;
//      if((maxfid-10)<0) fid0=0;
//      else fid0=maxfid-10;
        fid0=0;
        while(fid0<=maxfid && !check){
                check=1;
                for(int j=0; j<11 && check; ++j){
                        if(strcmp(big[fid0][j],fs[0][j])) check=0;
                }
                if(!check) ++fid0;
        }

        if(fid0>maxfid){
                ++maxfid;
                fid0=maxfid;
                for(int j=0; j<11; ++j){
                        strcpy(big[maxfid][j],fs[0][j]);
                }
                fprintf(field,"\n%d,",maxfid);
                for(int j=0; j<11; ++j) fprintf(field,"%s,",big[maxfid][j]);
//              for(int j=0; j<10; ++j) fs[0][j][0]=0;
```

```
    }

    check=0;
    fid1=0;
    while(fid1<=maxfid && !check){
        check=1;
        for(int j=0; j<10 && check; ++j){
            if(strcmp(big[fid1][j],fs[1][j])) check=0;
        }
        if(!check) ++fid1;
    }

    if(fid1>maxfid){
        ++maxfid;
        for(int j=0; j<11; ++j){
            strcpy(big[maxfid][j],fs[1][j]);
        }
        fprintf(field,"\n%d,",maxfid);
        for(int j=0; j<11; ++j) fprintf(field,"%s,",big[maxfid][j]);
//      for(int j=0; j<10; ++j) fs[1][j][0]=0;
    }
}

void parse();
void parse(){
    temp=line;
    thing = &temp;
    swap = strsep(thing, ",");
    if(!strcmp(swap,"id")) {
        strcpy(gid,*thing);
        pid = 0;
        inum = 0;
        fprintf(info,"\n%s,",gid);
//      printf("\ngid is %s",gid);
    }
    else if(!strcmp(swap,"info")){
        if(!((*thing)[0] == 't' && (*thing)[1] == 'r' && (*thing)[2] == 'a' && (*thing)[3]
== 'n' &&
            (*thing)[4] == 's' && (*thing)[5] == 'l' && (*thing)[6] == 'a' && (*thing)[7] ==
't' &&
            (*thing)[8] == 'o' && (*thing)[9] == 'r' && (*thing)[10] == ',')
          &&
          !((*thing)[0] == 'e' && (*thing)[1] == 'd' && (*thing)[2] == 'i' && (*thing)[3]
== 't' &&
            (*thing)[4] == 't' && (*thing)[5] == 'i' && (*thing)[6] == 'm' && (*thing)[7] ==
'e' &&
            (*thing)[8] == ',')
          &&
          !((*thing)[0] == 'i' && (*thing)[1] == 'n' && (*thing)[2] == 'p' && (*thing)[3]
== 'u' &&
            (*thing)[4] == 't')
          &&
          !((*thing)[0] == 's' && (*thing)[1] == 'c' && (*thing)[2] == 'o' && (*thing)[3]
== 'r' &&
            (*thing)[4] == 'e' && (*thing)[5] == 'r' && (*thing)[6] == ',')
          &&
          !((*thing)[0] == 'g' && (*thing)[1] == 'w' && (*thing)[2] == 'r' && (*thing)[3]
== 'b' &&
            (*thing)[4] == 'i' && (*thing)[5] == ',')
          &&
          !(inum!=2 && (*thing)[0] == 's' && (*thing)[1] == 'i' && (*thing)[2] == 't' &&
(*thing)[3] == 'e' &&
            (*thing)[4] == ',')
          &&
          !(inum!=1 && (*thing)[0] == 'h' && (*thing)[1] == 'o' && (*thing)[2] == 'm' &&
(*thing)[3] == 'e' &&
             (*thing)[4] == 't' && (*thing)[5] == 'e' && (*thing)[6] == 'a' && (*thing)[7] ==
'm' &&
             (*thing)[8] == ',')
          )
        {
            ++inum;
            swap = strsep(thing, ",");
            while(strcmp(swap,(force[inum-1])) != 0 && inum<29) {
                fprintf(info,"unknown,");
//                printf("%d inum should be %s, is %s\n",inum, force[inum-1], swap);
                ++inum;
            }
            fprintf(info,"%s,",*thing);
//            printf("OUTPUT: %d inum should be %s, is %s\n",inum, force[inum-1], swap);
            if(inum==1) strcpy(fs[1][10],*thing);
            if(inum==2) strcpy(fs[0][10],*thing);
        }
        //yay crazy hack
        else if((*thing)[0] == 'g' && (*thing)[1] == 'w' && (*thing)[2] == 'r' &&
(*thing)[3] == 'b' &&
                 (*thing)[4] == 'i' && (*thing)[5] == ',')
        {
            swap = strsep(thing, ",");
            if(strcmp(swap,(force[inum]))!=0) {
                fprintf(info,"unknown");
                ++inum;
            }
        }
    }
```

```
else if(!strcmp(swap,"start")){
	if(inum!=0 && inum!=27) printf("bad info on %d gid %s\n",inum,gid);
	//		if(inum==28) printf("%s ",gid);
	inum=0;
	togval=1;

	work = strsep(thing, ",");
	strsep(thing,",");
	swap = strsep(thing, ",");
	teamval = swap[0]-'0';
	swap = strsep(thing, ",");
	strcpy(fs[teamval][swap[0]-'0'],work);
	}
else if(!strcmp(swap,"play")){
	if(inum!=0 && inum!=27) printf("bad info on gid %s\n",gid);
	if(togval) {findfid(); togval=0;}
	inum=0;
	fprintf(play,"%s,%d,%d,%d,%s\n",gid,++pid,fid0,fid1,*(thing));
}
else if(!strcmp(swap,"sub")){
	if(inum!=0 && inum!=27) printf("bad info on gid %s\n",gid);
	inum=0;

	work = strsep(thing, ",");
	strsep(thing,",");
	swap=strsep(thing, ",");
	teamval=swap[0]-'0';
	swap=strsep(thing, ",");
	if(swap==NULL) printf("%s",gid);
	else strcpy(fs[teamval][swap[0]-'0'],work);
```

```
        findfid();


    }
    else if(!strcmp(swap,"com")){
        fprintf(com,"%s,%d,%s\n",gid,pid,*(thing));//don't add 1 to the pid, the com is
with the play
    }
    else if(!strcmp(swap,"data")){
        if(inum!=0 && inum!=27) printf("bad info on gid %s\n",gid);
        inum=0;
        fprintf(data,"%s,%s\n",gid,*(thing));
    }
    fflush(info);
    fflush(field);
    fflush(play);
//      fflush(sub);
    fflush(com);
    fflush(data);
    fflush(stdout);
}



int hasq(char*);
int hasq(char* x){//does the string have a quote
    int i=0, ans=0;
    while(x[i]!=0) if(x[i++]=='"') ++ans;
    return ans%2;
}

int main(){
    char teams[2][15][4] = {
    {{"CLE"},{"BAL"},{"BOS"},{"CAL"},{"CHA"},{"DET"},{"KCA"},{"MIL"},{"MIN"},
{"NYA"},{"OAK"},{"SEA"},{"TOR"},{"TEX"}},
        {{"PIT"},{"ATL"},{"CHN"},{"CIN"},{"HOU"},{"LAN"},{"MON"},{"NYN"},{"PHI"},{"SDN"},
{"SFN"},{"SLN"},{"Nul"},{"Nul"}}};

        for(int foo=0; foo<10; ++foo){

                char hack[11][7] =
{{"0.txt"},{"1.txt"},{"2.txt"},{"3.txt"},{"4.txt"},{"5.txt"},{"6.txt"},{"7.txt"},{"8.txt"},{"9.txt"},
{"10.txt"}};

                //              start = fopen(strcat("/Users/et/Desktop/BASEBALL/strat",hack[foo]), "w");

/*              fclose(info);

                fclose(field);

                fclose(play);

                fclose(sub);

                fclose(com);

*/

                swap = new char[100];

                swap[0]=0;

                printf("\nNEW SET");

                fflush(stdout);

                swap = new char[100];

                swap[0]=0;

                info = fopen(strcat(strcat(swap,"/Users/et/Desktop/DATA/info"),hack[foo]), "w");

                swap = new char[100];

                swap[0]=0;

                field = fopen(strcat(strcat(swap,"/Users/et/Desktop/DATA/field"),hack[foo]), "w");

                swap = new char[100];

                swap[0]=0;

                play = fopen(strcat(strcat(swap,"/Users/et/Desktop/DATA/play"),hack[foo]),"w");
//              swap = new char[100];
//              swap[0]=0;
//              sub = fopen(strcat(strcat(swap,"/Users/et/Desktop/DATA/sub"),hack[foo]),"w");
                swap = new char[100];

                swap[0]=0;

                com = fopen(strcat(strcat(swap,"/Users/et/Desktop/DATA/com"),hack[foo]),"w");

                swap = new char[100];
```

```
swap[0]=0;
data = fopen(strcat(strcat(swap,"/Users/et/Desktop/DATA/data"),hack[foo]),"w");
swap = new char[100];
swap[0]=0;

for(int j=0; j<2; ++j){
        for(int k = 1974+2*foo; k<=1992 && k<1976+2*foo; ++k){//1992
                for(int l=0; l<14-j*2; ++l){
                        if((l==11) && k<1977 && j==0) l+=2;

                        //                                      path[26]= '0' + k/1000;
                        //                                      path[27]= '0' + k/100%10;
                        path[29]= '0' + (k/10)%10;
                        path[30]= '0' + k%10;
                        if(j==0)
                                path[31] = 'a';
                        else path[31] = 'n';
                        if(k<1983+j && (k!= (1978+j)) && (k!=1981 || j!=1)){
                                path[34] = '0' + k/10%10;
                                path[35] = '0' + k%10;
                                path[36] = teams[j][l][0];
                                path[37] = teams[j][l][1];
                                path[38] = teams[j][l][2];
                                path[39] = '.';
                                path[40] = 'e';
                                path[41] = 'v';
                                if(j==0) path[42] = 'a';
                                else path[42] = 'n';
                                path[43]=0;
                        }
                        else{
```

```
                        path[34] = '0' + k/1000;
                        path[35] = '0' + k/100%10;
                        path[36] = '0' + k/10%10;
                        path[37] = '0' + k%10;
                        path[38] = teams[j][l][0];
                        path[39] = teams[j][l][1];
                        path[40] = teams[j][l][2];
                        path[41] = '.';
                        path[42] = 'e';
                        path[43] = 'v';
                        if(j==0) path[44] = 'a';
                        else path[44] = 'n';
                        path[45]=0;
                    }

                    //int fp = open (path,O_RDONLY);//the file being read
//                printf(path);
//                printf("\n");
                    printf("%7d",maxfid);
                    fflush(stdout);


                    FILE *fp = fopen(path, "r");
                    //gid[0]='0';
                    gid[0]=0;
                    do {
                        for(int i=0; i<100; ++i) imput[i]=0;
                        //                  line = new char[100];
                        //                  line = (char*) realloc( line, 100*sizeof
(char));
                        for(int i=0; i<100; ++i) line[i]=0;
                        fscanf(fp,"%s",imput);
                        strncat(line,imput,100);
                        if(hasq(imput)) {
                            do {
                                fscanf(fp,"%s",imput);
                                strcat(line," ");
                                strcat(line,imput);
                            } while(!hasq(imput));
                        }
                        if(line[0]=='i' && line[1]=='n' && line[2]=='f' && line[3]=='o' &&
line[4]==',' && line[5]=='i' &&
                          line[6]=='n' && line[7]=='p' && line[8]=='u' && line[9]=='t' &&
line[10]=='t' && line[11]=='i' &&
                          line[12]=='m' && line[13]=='e' && line[14]==',' ){
                            fscanf(fp,"%s",imput);
                            strcat(line," ");
                            strcat(line,imput);
                        }
                        if(line[0]=='i' && line[1]=='n' && line[2]=='f' && line[3]=='o' &&
line[4]==',' && line[5]=='e' &&
                          line[6]=='d' && line[7]=='i' && line[8]=='t' && line[9]=='t' &&
line[10]=='i' && line[11]=='m' &&
                          line[12]=='e' && line[13]==','){
                            fscanf(fp,"%s",imput);
                            strcat(line," ");
                            strcat(line,imput);
                        }

                        //                printf(line);
                        //                printf("\n");
                        parse();
                    } while(imput[0]!=0);
                    //              printf("\n\n");
                    fclose(fp);
                }
            }
        }
    }
    return 0;
}
```

Appendix B: Retrosheet Event File Sample

```
id,ATL197104090
version,1
info,inputprogvers,"version 7RS(19) of 07/07/92"
info,visteam,PIT
info,hometeam,ATL
info,date,1971/04/09
info,number,0
info,starttime,8:05PM
info,daynight,night
info,usedh,false
info,umphome,"Dick Stello"
info,ump1b,"Andy Olsen"
info,ump2b,"Augie Donatelli"
info,ump3b,"Bill Williams"
info,scorer,"Pirates"
info,inputter,"Alan Boodman"
info,inputtime,1996/02/01 12:11PM
info,howscored,unknown
info,pitches,none
info,temp,0
info,winddir,unknown
info,windspeed,-1
info,fieldcond,unknown
info,precip,unknown
info,sky,unknown
info,timeofgame,143
info,attendance,32734
info,wp,moosb101
info,lp,jarvp101
info,save,giusd101
info,gwrbi,
start,mazeb101,"Bill Mazeroski",0,1,4
start,hebnr001,"Richie Hebner",0,2,5
```

```
start,clemr101,"Roberto Clemente",0,3,9
start,starw101,"Willie Stargell",0,4,7
start,oliva001,"Al Oliver",0,5,8
start,robeb101,"Bob Robertson",0,6,3
start,sangm101,"Manny Sanguillen",0,7,2
start,hernj102,"Jackie Hernandez",0,8,6
start,moosb101,"Bob Moose",0,9,1
start,jacks102,"Sonny Jackson",1,1,8
start,garrr101,"Ralph Garr",1,2,7
start,aaroh101,"Hank Aaron",1,3,9
start,cepeo101,"Orlando Cepeda",1,4,3
start,kingh102,"Hal King",1,5,2
start,millf105,"Felix Millan",1,6,4
start,boyec102,"Clete Boyer",1,7,5
start,perem101,"Marty Perez",1,8,6
start,jarvp101,"Pat Jarvis",1,9,1
play,1,0,mazeb101,??,,W
play,1,0,hebnr001,??,,5
play,1,0,clemr101,??,,D.1-3
play,1,0,starw101,??,,W
play,1,0,oliva001,??,,46(1)3/GDP
play,1,1,jacks102,??,,7
play,1,1,garrr101,??,,S
play,1,1,aaroh101,??,,PO1(13)
play,1,1,aaroh101,??,,4/L
play,2,0,robeb101,??,,8
play,2,0,sangm101,??,,31
play,2,0,hernj102,??,,S
play,2,0,moosb101,??,,13
play,2,1,cepeo101,??,,D
play,2,1,kingh102,??,,S.2-3
play,2,1,millf105,??,,13.1-2
play,2,1,boyec102,??,,IW
play,2,1,perem101,??,,64(1)3/GDP
play,3,0,mazeb101,??,,63
play,3,0,hebnr001,??,,3/G
```

```
play,3,0,clemr101,??,,63
play,3,1,jarvp101,??,,53
play,3,1,jacks102,??,,9
play,3,1,garrr101,??,,S
play,3,1,aaroh101,??,,63
play,4,0,starw101,??,,S
play,4,0,oliva001,??,,E3.1-2
play,4,0,robeb101,??,,64(1)/FO.2-3
play,4,0,sangm101,??,,S.1-2;3-H(UR)
play,4,0,hernj102,??,,K
play,4,0,moosb101,??,,63
play,4,1,cepeo101,??,,53
play,4,1,kingh102,??,,7
play,4,1,millf105,??,,S
play,4,1,boyec102,??,,WP.1-2
play,4,1,boyec102,??,,53
play,5,0,mazeb101,??,,S
play,5,0,hebnr001,??,,S.1-2#
play,5,0,clemr101,??,,S.1-2;2-H
play,5,0,starw101,??,,NP
sub,barbs102,"Steve Barber",1,9,1
play,5,0,starw101,??,,16(1)/FO.2-3
play,5,0,oliva001,??,,S.1-2;3-H
play,5,0,robeb101,??,,3
play,5,0,sangm101,??,,S.1-2;2-3
play,5,0,hernj102,??,,D.1-3;2-H;3-H
play,5,0,moosb101,??,,31
play,5,1,perem101,??,,43
play,5,1,barbs102,??,,NP
sub,staem101,"Marv Staehle",1,9,11
play,5,1,staem101,??,,K
play,5,1,jacks102,??,,43
play,6,0,mazeb101,??,,NP
sub,herbr102,"Ron Herbel",1,9,1
play,6,0,mazeb101,??,,9
play,6,0,hebnr001,??,,7
```

```
play,6,0,clemr101,??,,S
play,6,0,starw101,??,,HR.1-H
play,6,0,oliva001,??,,HP
play,6,0,robeb101,??,,CS2(24)
play,6,1,garrr101,??,,S
play,6,1,aaroh101,??,,5
play,6,1,cepeo101,??,,6/L
play,6,1,kingh102,??,,4
play,7,0,robeb101,??,,8
play,7,0,sangm101,??,,43
play,7,0,hernj102,??,,63
play,7,1,millf105,??,,8
play,7,1,boyec102,??,,13
play,7,1,perem101,??,,3/G
play,8,0,moosb101,??,,63
play,8,0,mazeb101,??,,S
play,8,0,hebnr001,??,,8
play,8,0,clemr101,??,,64(1)/FO
play,8,1,herbr102,??,,NP
sub,wille101,"Earl Williams",1,9,11
play,8,1,wille101,??,,S
play,8,1,jacks102,??,,S.1-3
play,8,1,garrr101,??,,36(1)/FO.3-H
play,8,1,aaroh101,??,,64(1)/FO
play,8,1,cepeo101,??,,5
play,9,0,starw101,??,,NP
sub,baked002,"Dusty Baker",1,9,9
play,9,0,starw101,??,,NP
sub,kellt102,"Tom Kelley",1,3,1
play,9,0,starw101,??,,S
play,9,0,oliva001,??,,W.1-2
play,9,0,robeb101,??,,S.1-2;2-H
play,9,0,sangm101,??,,9.2-3
play,9,0,hernj102,??,,6(1)3/GDP
com,"May have been an LDP."
play,9,1,kingh102,??,,S
```

```
play,9,1,millf105,??,,1
play,9,1,boyec102,??,,S.1-3#
play,9,1,perem101,??,,7
play,9,1,baked002,??,,D.1-3;3-H
play,9,1,jacks102,??,,NP
sub,giusd101,"Dave Giusti",0,9,1
play,9,1,jacks102,??,,13
data,er,moosb101,2
data,er,giusd101,0
data,er,jarvp101,3
data,er,barbs102,1
data,er,herbr102,2
data,er,kellt102,1
```

Appendix C: Handout at 4/16 SABR Meeting

This is the text of a handout made available during the second SABR meeting attended by the project group. Pat Malloy participated

Statistical Baseball Analysis Database

Project Members:

Pat Malloy - pmalloy@wpi.edu

Kevin Munn - munnk@wpi.edu

Rory Fuller - rgf@wpi.edu

Ethan Thompson - et@wpi.edu

- What did we do?

We have taken the data from 1974 though 1992 from RetroSheet and converted it into an Access database. In this format is is much easier

to manipulate the data using existing database techniques.

- Why did we do this?

We are an IQP (Interdisciplinary Qualifying Project) group developing this database as the main part of the project. Being allowed to develop our own projects let us pick something interesting while still satisfying the school requirement.

- What are the key points to the database?

Defensive statistics:

Have the players in the field easily accessible

Have the sector from Project Scoresheet (when available)

Have the pitcher

For all data in the Project Scoresheet era, have (by and large) count,

pitch order

Offensive Statistics:

Potential for a vector model for hitting

Adjust hits based on quality of defense

- Does it work?

We are, of course biased, but we believe it works beautifully, and we already have other groups at WPI planning on using it as their basis for more advanced statistical analyses.

- How do I get this database?

It isn't ready for distribution yet, though we expect it done by early summer. Please either e-mail one of us, or preferably, sign up with us after the talk, and we will let you know when it is ready. If at all possible we will be distributing primarily though downloads off the internet.

References

[1] Hall, Alvin (1989). *Cooperston Symposium on Baseball and American Culture*. 1st Ed. Westport, CT: Meckler in association with State Universityof New York College at Oneonta.

[2] Lewis, Michael (2003). *Moneyball: the Art of Winning an Unfair Game*. 1st Ed. New York: W. W. Norton.

[3] Major League Baseball Player's Association (2005). "Frequently Asked Questions." Retrieved April, 2005. http://mlbplayers.mlb.com/NASApp/mlb/pa/info/faq.jsp#average

[4] PBS (2001). "Wonders of the World Databank: Sears Tower." Retrieved April, 2005. http://www.pbs.org/wgbh/buildingbig/wonder/structure/sears_tower.html

[5] ITMS Sports (2002). May 2002 News Archive. Retrieved February, 2005. http://www.epdtour.de/news/narchive/may02.htm

[6] Sports Illustrated News (2005). "Baseball ticket prices up more than 6 percent, highest rise since 2000." Retrieved April, 2005. http://sportsillustrated.cnn.com/2005/baseball/mlb/wires/04/04/2010.ap.bbo.ticket.prices.0390/

[7] CIA (2005). World Factbook Homepage. Retrieved May, 2005. http://www.cia.gov/cia/publications/factbook/index.html

[8] MovieWeb (2005). "All-Time 100 Top-Grossing Films (US)." Retrieved April, 2005. http://movieweb.com/movies/box_office/alltime.php

[9] Japanese Baseball (2005). Index of Professional Japanese teams, circa 2005. Retrieved April, 2005. http://japanesebaseball.com/teams/index.jsp

[10] Gordon Library Online Database and Journal Index (2005). Homepage. Retrieved March, 2005. http://www.wpi.edu/Academics/Library/Collections/Databases/

[11] Baseball Prospectus (2005). Homepage. Retrieved January, 2005. http://www.baseballprospectus.com/

[12] Amazon Online Bookstore (2005). Homepage. Retrieved April, 2005. www.amazon.com

[13] Elias Sports Bureau (2005). Homepage. Retrieved January, 2005. www.esb.com

[14] Retrosheet (2005). Homepage. Retrieved January, 2005. www.retrosheet.org

[15] Schell, Michael J (2005). Sample chapter from *Baseball's All-Time Best Hitters: How Statistics Can Level the Playing Field.* Retrieved April, 2005. http://www.pupress.princeton.edu/chapters/i6550.html

[16] Microsoft Encarta Reference Library Product Page (2005). product Homepage. Retrieved May, 2005. http://www.microsoft.com/products/encarta/ProductDetails.aspx?pid=004

[17] Kenda, Rich (2001). *USA Today Baseball Weekly*. “Where do those crazy statistics come from?” Retrieved April, 2005. http://www.usatoday.com/sports/bbw/2001-12-12/2001-12-12-clubhouse.htm

[18] STATS, Inc. (2005). Homepage. Retrieved January, 2005. http://www.stats.com/

[19] Baseball Reference (2005). Homepage. Retrieved January, 2005. www.baseball-reference.com

[20] Bball Sports Online Interactive Historical Sports Statistics Databases (2005). Homepage. Retrieved January, 2005. www.bballsports.com

[21] The Baseball Archive (2005). Homepage. Retrieved January, 2005. http://baseball1.com

[22] 7-Zip (2005). Homepage for free extractor program. Retrieved May, 2005. http://www.7-zip.org/

[23] Mathematical Association of America (2005). Homepage. Retrieved May, 2005. http://www.maa.org/

[24] Albert, Jim (2005). *Teaching Statistics Using Baseball* online resource page, Retrieved May, 2005. http://personal.bgsu.edu/~albert/teachball.htm

[25] Albert, Jim (2001). *Curve Ball: Baseball, Statistics, and the Role of Chance in the Game*. 1st Ed. New York: Copernicus.

[26] Hardball Times (2005). “Win Shares for MVP.” Retrieved February, 2005. http://www.hardballtimes.com/main/article/win-shares-for-mvp

[27] Baseball Graphs (2005). Baseball Graph Details, “How are win shares calculated?” Retrieved March, 2005. http://www.baseballgraphs.com/details.html#sharecalc

[28] Baseball Musings (2005). Homepage. Retrieved April, 2005. http://www.baseballmusings.com/

[29] Humphreys, Michael (2005). Hardball Times. Defensive Regression Analysis: Complete Series. Retrieved April, 2005. http://www.hardballtimes.com/main/article/defensive-regression-analysis-complete-series/