

# **The History of Total Baseball and Pete Palmer's Baseball Databases**

*By Pete Palmer*

## **In The Beginning: The Barnes Official Encyclopedia of Baseball**

I started my career as a baseball historian and statistician while editing the 7th through 10th editions (1974–1979) of the old Turkin-Thompson encyclopedia, published by A.S. Barnes. Julien Yoseloff gave me the job in the first place, after a few years of prodding. My first step was to convert all of their data to computer files, including names, birth and death info, years played, teams, positions, games, and batting averages for players plus wins and losses for pitchers. I also did a revision of Thompson's All-Time Rosters book, which added games by position for modern players. The next step was to add at bats, hits, and walks plus hit-by-pitch, innings pitched, and earned run averages in order to calculate linear weight values for all players. These linear weights ended up being the core of the analysis for *The Hidden Game of Baseball*, which I co-authored with John Thorn in 1984.

This initial data compilation was all done on IBM 80-column punch cards and then read into a mainframe computer, which I was able to use during off hours at my job at Raytheon. I had previously punched out all of the 1969 and 1970 batting and pitching data while doing an analysis of the Mills brothers' Player Win Averages, and I had found it not too difficult. Therefore, throughout this period I was adding complete batting, pitching, and fielding data from the annual official guides. By the time *The Hidden Game* came out, I had already gotten back to 1925, and I completed punching out the 1871–1924 stats by 1988 for the first edition of *Total Baseball*. By then I had dumped all of my mainframe data onto floppy disks and converted to using a PC.

## **Biographical, Manager, Coach, and Umpire Data**

The biographical data for my original database was compiled from the Turkin-Thompson book. The legendary Bill Haber helped by updating my biographical information in the 1970s since it had not been well maintained after S.C. Thompson died. Cliff Kachline made available the newsletters of his bio group from 1971-onward. Kachline's project became the SABR Biographical Committee, which has become a primary source of updated biographical data. The Biographical Committee was later headed by Rich Topp and is now chaired by Bill Carle. When I find missing biographical information or identify a mistake in published biographical info, I send it to the Biographical Committee, which distributes the new information via its newsletter.

Rich Malatzky and Peter Morris have been particularly prolific in contributing to this biographical research. Fred Lenger has investigated every single place of birth or death in the database, trying to identify the exact location, and has come up with many

corrections. The manager data came from a SABR research project headed by Bob Tiemann and Rich Topp, while the coach data was largely from Bob Hoie. Umpire data came originally from S.C. Thompson, but was expanded by Larry Gerlach and John Schwartz.

### **Making the Official Stats Complete**

Prior to 1941, the official stats in the annual guides were incomplete for players who appeared in fewer than 10 games, so I had to use the official microfilm to get complete data for those players. I had been using the AL microfilm that was available at the league office in Boston when Joe Cronin was president; luckily, I was able to get a copy of that microfilm through Tom Monahan there before the AL office was moved to New York City. For the NL microfilm, I had to go through the Hall of Fame Library in Cooperstown, where Cliff Kachline and Tom Heitz were helpful. Leonard Gettelson also helped fill in some of the missing NL data for the early twentieth century. Recently, Steve Gietschier of The Sporting News allowed me to obtain a copy of their NL microfilm to fill in the years I was missing.

Official league stats only go back to 1902 in the National League and 1905 in the American League. As a result of the work that David Neft and his associates at ICI did in the late 1960s for the original edition of the Macmillan Baseball Encyclopedia, the Hall had computer printouts covering the earlier years, except for 1876–1890 in the NL

### **The Pioneering Work of John Tattersall, Clarence Dow, Michael Stagno, Bob Tiemann, and Bob McConnell**

For those 15 years in the nineteenth century, the primary source was John Tattersall's collection, which was purchased by SABR upon his death and maintained by Bob McConnell. Tattersall had years earlier obtained a large set of old newspapers, and thus had the raw information to complete the published official stats—which in those years contained only games, at bats, runs, and hits. Tattersall added in extra-base hits, walks, and strikeouts from these box scores and compiled runs batted in from the newspaper game accounts. He also obtained data for players who were traded or who played in fewer than 15 games in a season from these box scores. For some years, Tattersall used data compiled by Clarence Dow—one of the all-time great statisticians—as published in the Boston Globe. Dow, who actually played a game for the Boston Unions in 1884, kept detailed stats and is the only reason we have batter strikeouts up through 1896. Unfortunately, Dow died prematurely in 1893, and his work was continued only briefly after his demise.

Part of my collection of source material are printed or microfilmed back-up data for box scores from Bob Tiemann for 1871–1875, from the Tattersall scrapbooks from 1876–

1882, from *Sporting Life* from 1883–1916, from *Sporting News* from 1886–1990, from *Baseball Weekly/Sports Weekly* from 1991–2002, and from Internet sources since then. McConnell and I made copies of Tattersall's stats sheets for my use, and he provided all of the corrections he and Tattersall had made while compiling what became the SABR Home Run Log. McConnell also provided data from his extensive research into switch-hitters.

Recently, the Hall of Fame has made available to the public microfilm copies of the ICI printouts, the Tattersall box score collection, and the official NL and AL microfilms. National Association data came originally from box scores collected by Michael Stagno and purchased by SABR. Bob Tiemann, ably assisted by Bob Richardson, enhanced the Stagno collection considerably, including finding box scores for some of the handful of missing games. Tiemann also did extensive research on nineteenth century attendance. His notebooks, computerized by Arnie Braunstein, were the basis for the Retrosheet.org game logs, produced by Tom Ruane. Ruane added detailed game data from Retrosheet's own play-by-play data up through 1983, from the former Project Scoresheet play-by-play data from 1984–1990, and from Gary Gillette's Baseball Workshop play-by-play data for 1991–1996. As part of this effort, Ruane and I resolved all differences in scores and locations for every game since 1900. Joe Simenic provided the park Cleveland used for all games in the 1932–1946 period where the Indians split play between League Park and Cleveland Municipal Stadium.

### **Filling in the Gaps**

Despite this huge amount of research, there were still holes in the available data. One of the largest was in batters hit by pitch stats, which started in 1884 for the American Association and in 1887 for the NL but was not kept officially by the NL until 1917 by the AL till 1920. I did the NL from 1887–1896 from Boston Globe box scores at the Boston Public Library, along with 1890 Players League and the 1891 AA. I used Tattersall's data for 1884–1889 AA and did the 1890 AA from *Sporting Life*. Alex Haas had previously done NL HBP from 1909–1916 and AL for 1909–1919 from New York Times box scores (the first years they were carried by the Times). A group SABR effort directed by John Schwartz filled in 1897–1908 from local newspapers for the still missing years. I compiled pitcher hit batsmen, wild pitches, and balks from *Sporting Life* up till 1903 in the NL and 1908 in the AL, when they were first kept officially. I used the annual Spalding Guides for batter sacrifice hits from 1893 (when they were first counted) until 1903 in the NL and 1905 in the AL. Neil Munro helped with traded players and players who appeared in fewer than 15 games: None of these items were in the Macmillan printouts.

Because Munro was working on his own career database and we were helping each other with some items, Munro also contributed games finished (which were not kept officially until 1912 in the AL and 1919 in the NL). The NL kept at bats versus pitchers up till 1912, when the league shifted to total batters faced. However, the NL did not include

sacrifice hits allowed for 1912, making it impossible to calculate at bats for that season. With help from many SABR members—including Joe Dittmar, Walt Wilson, Ralph Horton, Paul Doherty, Ed Luteran, Neil Traven, Herb Goldman and Tom Chase—I filled in this data from newspaper game accounts. The AL did not start sacrifice hits allowed by pitchers until 1921; thus it is impossible to calculate BFP before then. The league did, however, mistakenly record BFP instead of at bats in a great many games, as did the NL before 1912. Some of these mistakes have been found; some haven't—I discovered that the erroneous data was typically concentrated in certain cities for parts of a particular season, though no one else has ever attempted to find and fix this problem.

I compiled games played in left, center, and right field for all outfielders from box scores for 1876–1890 NL, from the ICI computer sheets where they existed, and from the AL microfilm. Bill Deane and the Hall of Fame research staff filled in NL data from their microfilm.

Caught stealing data has been particularly elusive. Official data began in 1920, but the AL dropped it for one year in 1927 and the NL didn't keep track of it for 25 years from 1926–1950. Another SABR research effort filled caught stealing in the AL for 1927 from newspaper accounts; this group included Bob Richardson, Paul Doherty, Walt Wilson, Rich Topp, Don Luce, Bill Shlensky, Mike Weddell, Walter LeConte, Tom Howell, Tim Swindle, and Bob Davids. Regrettably, the NL gap remains unfilled to this day. Ernie Lanigan kept data unofficially from 1912–1919, but only about half of Lanigan's data has been found so far in articles in various newspapers. With the help of Bob Davids, I have found 1913 and 1915–1916 NL data as well as 1914–1916 AL data, though the 1916 data is only for players with 20 or more stolen bases. Jim Weigand looked into stolen base/caught stealing figures and found many errors in the AL, particularly in the 1920s.

In 1954, the first year of the sacrifice fly rule, the AL did not break down sacrifices allowed by pitchers into sac bunts and sac flies. Fortunately, I was able to separate these two stats from Sporting News and New York Times game accounts.

Another problem area was AL pitching stats for 1901–1919, which were particularly error-prone even when published in the Macmillan encyclopedia (which took its data from the official records). Frank Williams went over this very data carefully and came up with revised figures for games, games started, complete games, shutouts, saves, wins, losses, relief wins, and relief losses. Williams' corrected data was later used by Macmillan and by David Neft in *The Sports Encyclopedia Baseball* as well as by me.

Joe Wayman investigated every shutout and compiled revised data for all seasons from 1876 to date. He identified combined shutouts (which were often credited to a single pitcher in the early years) and shutouts of less than 9 innings (which sometimes were not awarded). John O'Malley did a great deal of nineteenth century research, particularly about the career of Tim Keefe.

Incidental research at various times has been done in Boston by Bob Richardson, in Chicago by Walt Wilson, in Cleveland by Joe Simenic, in New York by Herb Goldman,

in St. Louis by Keith Carlson, and at the Hall of Fame by Bill Deane and Eric Enders. A great number of other people have contributed over the years, including Greg Beston, Tim Cashion, Dan Dischley, Cam Gibson, Ray Gonzalez, Ed Hartig, Ron Liebman, Trent McCotter, Wayne McElreavy, and Lyle Spatz.

### **The Best Laid Plans**

Walt Wilson researched games started prior to their adoption as an official stat in 1926 in the AL and 1938 in the NL. The ICI-Macmillan data had many errors in NL games started due to the format of the official data, which could be very confusing if someone was trying to figure out games started. While it is a long story about a single stat, it is a perfect illustration of how difficult it is to get things right sometimes, as well as how what seems like a perfectly reasonable procedure can introduce unexpected errors in the stats.

In this case, the official sheets had a comment column that allowed a statistician to enter “relieved Smith,” “relieved by Smith,” or both. So, one could figure out that if this column was blank, it was a complete game. Otherwise, it was a start if the comment was “relieved by,” a game finished if the pitcher “relieved” someone else but was not himself replaced, and an incomplete game if both types of comments were present. However, these notes were not well kept; even worse, they often had ditto marks that could mean different things. If the top line said “relieved by,” the second line might have a ditto that applied only to the “relieved” part of the previous comment; in some cases, there were two sets of ditto marks if it pertained to both parts of the previous comment. In other cases, the ditto marks wouldn’t mean the same thing. Macmillan tried to use these to avoid the extra work of going through the box scores; even though they did a good job, Walt Wilson still found more than 100 corrections.

### **Putting It All Together: The Publication of Total Baseball**

Over the years, I have found and corrected many thousands of errors in the published official statistics. Many others have contributed to this lengthy and complicated process as well, as noted in this history. A great number of errors were discovered by the simple process of comparing the sum of all players on a team to the team total. NL sums of player batting and fielding statistics agreed with team totals starting in 1910, so this was in pretty good shape from then on. AL player sums did not reliably match their team totals until 1935. The NL still had many pitching stat errors up until the mid-1940s, while the AL did not get complete agreement between individual pitcher sums and team totals until it started using computers in 1973. Bill Carr made up a complete list of differences between my data and the Macmillan data for 1876–1900; I was able to resolve each one, some which resulted in my correcting a few errors in my database.

After more than two decades of hard work, this database was presented in the first edition Total Baseball in 1989 and was at the core of all eight editions of Total Baseball. The version of that database used in the 3rd edition of TB in 1993 was also published on CD by Creative Multimedia Corporation (CMC) of Portland, Oregon. It was then reverse-engineered off the CD by Sean Lahman and published on the Internet in the mid-1990s.

Tom Ruane later compared the CMC Total Baseball CD data to the Stats CD, which was based on Neil Munro's research. Ruane came up with a couple of thousand differences, every one of which I was able to resolve. Ed Yerha did the same for biographical data with the same results.

Thus, the database that I compiled for Total Baseball is the source of almost all online databases today, including those at prominent Websites like Baseball1.com, Baseball-Reference.com, Retrosheet.org, and Baseball-Databank.org. Licensed versions of that Total Baseball database appear as well on MLB.com and in Tom Tippett's Diamond Mind baseball game.

### **Birth of the New Baseball Encyclopedia**

My current database was created in the past few years from the same original sources in essentially the same way. While it was a large task, I had the huge advantage of having kept complete records of corrections and revisions made by me, by SABR, and by many others over the years. Furthermore, I had written many programs over the years to compile and test the data and was able to use them as well. So it wasn't exactly like reinventing the wheel—it was more like moving from one company to another and using 30-plus years of experience to build a better product. Along the way, I found and corrected many small typographical mistakes that I hadn't noticed previously.

My new database was the foundation for the career registers in the 2004 edition of Barnes & Noble's Baseball Encyclopedia that I co-edited with Gary Gillette. The new edition—now called The 2005 ESPN Baseball Encyclopedia—is being published this spring by Sterling Publishing, a subsidiary of Barnes & Noble.

### **A Final Note About Cap Anson**

It was while compiling the data for my new database that I discovered the ill-advised changes Tattersall had made in Cap Anson's hit totals for 1889. As we explained in the introduction to our new encyclopedia, Anson's published career hits totals in the past 35 years have varied from encyclopedia to encyclopedia—sometimes from edition to edition as well.

After careful consideration and a lot of investigation, Gillette and I agreed to restore Anson's 1889 stats to the official totals in The 2004 Baseball Encyclopedia. Tattersall also made similar changes to several other White Stockings batters in 1889, which we also restored.

<http://www.sabr.org/sabr.cfm?a=cms,c,67,35>

Created On: 02.12.05