September 29, 2011

New type of file on Retrosheet website

Overview by Dave Smith

Below is a detailed description of a major new effort by Retrosheet, one that has never been undertaken by any other group or company or by Major League Baseball itself. We have done a detailed examination of the game by game totals for batters, pitchers and fielders to check for inconsistencies in the official records. We have recently published the first set of results from this examination for several seasons, following the detailed methodology laid out below.

What are the official records in baseball? This turns out to be less obvious than most people understand. Leaving aside the modern, instantaneous computerized recording of events, the historical pattern had two basic steps. First, the official scorer at the game completed a standardized form with the totals for each player in all the categories required in the rule book. These have varied over the years, but there are around 25 for batters (including fielding data) and about the same for pitchers. Second, this form was sent to the American or National League office where the numbers were transcribed onto large ledger pages, a separate one for each player and pitcher. At the end of the season, the columns were totaled for each player and they formed the basis of the summaries in the annual guides (Spalding, Reach, The Sporting News). Examples of these handwritten pages may be seen for a 1927 batter at http://www.retrosheet.org/Ruth1927Page5.pdf and a 1934 pitcher at http://www.retrosheet.org/Dean1934Page1.pdf .

We must always remember what "official" means. The literal meaning is "of the office," not "guaranteed to be accurate." That is, a designation of "official" is a statement of authority, not of reliability. The ledger pages have long since been transferred to the Hall of Fame Library in Cooperstown where they have been microfilmed. Retrosheet is fortunate to have been able to purchase a copy of these microfilms (nearly 200 reels) and they were the basis of the work we have just released. Of course, all of this information is now transmitted from the press box to the MLB offices in "real time", so the issue is primarily for games played prior to 1980.

Most people have never seen these images of the official daily pages and the concept of "official" is therefore usually associated with season or career totals. There have been many disputes and controversies about these aggregate totals, but in reality there is no single official set of yearly or career totals, even though various encyclopedias have been designated as "official" over the years. MLB tends to defer to the Elias Sports Bureau

when disagreements arise, but Elias does not publish its information either in print or electronic versions, with one consequence being that resolution of differences is difficult.

The distinction between "official" and "accurate" may seem odd in 2011, but 100 years ago, it was a very serious issue. The late Leonard Koppett, 1992 recipient of the J.G. Taylor Spink award from the Baseball Writers Association of America, urged caution in these matters, noting that in the early 20th century there were "dueling statistics" (his phrase to me) and that everyone was better served by having one set agreed upon by everyone. It was this conversation that led me to appreciate the powerful significance of "from the office." I wrote an article for Memories and Dreams, the Hall of Fame publication, on this topic in the spring of 2011. A longer version of that story is at http://www.retrosheet.org/BaseballRecords.pdf .

How does all of this relate to Retrosheet's recent release? The approach was to examine the daily official totals from the microfilm for each game and to check for internal consistency. For example, as noted below, the number of strikeouts charged to one team's batters must equal the number of strikeouts credited to the opposing team's pitchers. Other types of checks were done as well (again, see below for details), and thousands of inconsistencies have been identified and carefully documented. At Retrosheet, we have traditionally referred to these as "discrepancies" and not as "errors" since the latter term has a great potential to be inflammatory, with the likely result that antagonistic relations will obscure the overall effort. However, many of the differences that were found truly are errors and it is time that we note them that way.

Details of the Proofing Process by Tom Ruane

In light of the recent release of Retrosheet's discrepancy files, I thought it might be a good idea to discuss their history and explain how the first set will differ from subsequent releases.

Back in July of 2004, Retrosheet announced the formation of the Box Score Project, an effort designed to generate computerized box score event files (a format devised by Ted Turocy and myself years earlier while we were working on the 1911 NL) for the years prior to those covered by our event files.

This effort had three main components:

1) digitizing lineup, line score and any other information not in the dailies that would be needed to generate box scores,

2) digitizing the Hall of Fame player, pitcher and team dailies and,

3) combining the results of the first two steps into proofed box score event files.

We started with a call for volunteers that summer and within six years had a complete set of box score event files from 1920

to 1949.

Now our volunteers are meticulous and dedicated people, but they are not infallible, so we needed to ensure that all of the work we received was thoroughly proofed in order to minimize the number of errors we released. So we did a series of sanity checks on all the digitized files we received. Some of the checks we did were obvious. We ensured, for example, that:

1) the sum of each batter or pitcher's dailies matched their official seasonal totals,

2) each team had a player at each position at both the beginning and end of each game,

3) the sum of a team batters and fielders' in each game matched the applicable value in the team dailies and,

4) the sum of a team batters' data matched the sum of the opposing team's pitchers' data.

And some of our checks were not as obvious. For example, ensuring that:

1) each team's plate appearances by lineup position made sense, unless there was a documented case of the team batting out of order during the game and,

2) no one player scored and drove in all of his team's runs without being credited with the same number of home runs.

There were dozens of different checks, but all of them were designed to eliminate data entry errors on the part of our volunteers and allow us to produce box scores on our site that made sense. What we found, however, was that once all of the data entries were corrected, the official Hall of Fame dailies still failed many of these checks. In short, each league/season contained hundred of instances where the official dailies had to be wrong.

The most common error involved having incorrect defensive positions marked so that, for example, a team officially had two second basemen in a game and no shortstop. But there were a wide variety of other apparent official mistakes and all of them had to be investigated and resolved. In some cases,

we couldn't determine what actually happened and have left anomalies in the box scores. But most of the time, we found evidence to support a version of the game that corrected these errors. To be sure, mistakes still exist and I'm sure there are occasions where we have compounded the original error by a mistaken "correction," but these discrepancy files are the first step to both documenting possible errors in the official record as well as helping to bring to light our mistakes. This is certainly not the final word on the subject; rather, we hope it is the start of a dialogue.

The format of the discrepancy files are described in detail at http://www.retrosheet.org/Discrepancy%20File%20Format.pdf but they are currently displayed in three areas on our web-site:

1) Each team page, if there are discrepancies associated with that team, contains a link to a page describing them,

2) Each player page, if there are discrepancies associated with that player, contains a link to a page describing them,

3) Each game (box score) page, if there are discrepancies associated with that game, contains a link to a page describing them,

Earlier I had hinted that the first set of discrepancy files will differ from subsequent ones. What I meant was that this release contains discrepancies between our box score event files (created from Official Dailies) and the official statistics (or least Pete Palmer's statistical DB, which is even better than the official statistics). As a result, almost all of the discrepancies in these files involve cases where the official version has to be incorrect. No team can play a game with two second-basemen and no shortstop. No player can go 2-2 with two strikeouts, or have more home runs than RBIs, and so on. We may be wrong in how we resolved the problem, but we know that the official version can't be right.

Subsequent releases will also have discrepancies like this, but in addition they will include differences between the Hall of Fame dailies and the statistics derived from our play-by-play files. In these cases, both views of the game may be equally plausible on the surface; we simply believe that the play-by-play file is correct. In the next release, we will be including the discrepancy file for the 1922 NL so this might be a good example. There are a total of 1223 discrepancies in that file. 356 of them are differences between the box score files and the official dailes (in other words, instances where we can be confident there IS an error in the official view of things). But the other 867 are not as clear-cut. We have invesigated the games involved and think our version is the correct one, but further research may not uphold this view.

One final word: almost all of the initial generation of these discrepancy files was done by a single Retrosheet volunteer. If you have been reading the credits on the website or in our release notes over the last decade or so, Dave Lamoureaux's name should be familiar to you by now. He was responsible for a lion share of both of the first two digitization steps listed above (the lineup entry and the dailies), and he was a one-man show when it came to taking lists of statistical discrepancies and turning it into a game-by-game description of all of our differences. Much like the box scores from this era, these discrepancies files would not exist without thousands of hours of his work.