The RetroFile Project

Ted Turocy (arbiter@nwu.edu)

June 23, 2000

- What is RetroFile?
  - RetroFile is a small group of Retrophiles who have been thinking about the issues involved in extending the Retrosheet file format to permit representation of boxscore-level information and "partial accounts", particularly from older seasons.
  - Some of the issues involved:
    - In working from newspaper accounts, we often miss innings where neither side scored, and also lack later innings played after the final edition of the newspaper
    - In some cities, no (or few) play-by-play newspaper accounts are extant, meaning there are likely to be gaps in our collection for some time
    - The information included in boxscores has varied widely over the years, and across sources.
- An (informal) proposal
  - For games for which we do not have complete play-by-play, we propose to extend the file format via new game-level data lines
  - Example: Batting lines

    `data,bline,player,slot,seq,ab,r,h,2b,3b,hr,rbi,`

    `bb,ibb,so,gdp,hp,sh,sf,sb,cs`

    `data,bline,wagnh101,4,1,4,1,1,1,0,0,,0,,1,,0,0,0,0,`
  - Similar lines for pitching, fielding, pinch-hitting, and pinch-running performance, and overall team-level totals (e.g., double plays, left on base)
  - Also, optional lines detailing participants in double or triple plays, and modern "expanded" boxscore information such as batteries for stolen bases and caught stealing, and pitchers victimized by home runs.
  - Linescores will be represented by

    `data,linescore,team,inning,runs`

    For example,

    `data,linescore,0,1,1`

    for a visiting team scoring one run in the top of the first

  - In the case of partial play-by-play, we propose that a complete half-inning is the smallest fragment we will input. (Incomplete half-innings seem to cause more complications than they are worth.) The complete entry for half-innings will replace the linescore entry (that is, we can think of linescore entries as "placeholders" for missing half innings of play-by-play)

- Current status
  - Pilot project: working on the 1911 National League
  - Plan of attack:
    * Computerize the NL dailies (Tom Ruane) and *New York Times* boxscores (Ted Turocy), and merge the datasets to look for discrepancies
    * Produce a complete set of boxscore files for the season
    * Reconcile existing input games against these totals, and process partial accounts
    * Compare our data against other sources in the case of outstanding discrepancies

- A Website proposal
  - You can check out the current status of the project at
    `http://oskar.kellogg.nwu.edu/boxscore`
  - This site allows one to interactively browse boxscores, player daily performance, splits, etc.
  - This concept seems to be a natural fit as an enhancement to the Retrosheet site:
    * Convenient interface to the data for casual visitors
    * For older seasons, the boxscore-level data will be a natural byproduct of the process described above
    * Valuable tool for researchers
    * A "hook" for recruitment, and increasing awareness among people who might come across missing games