Major League Baseball Box Scores on ESPN.com

#standards #classification #authority #staticproperties #bias #SportsML #robo-journalism

**Overview.** Having played baseball from tee-ball through college, my relationship with the box score is quite personal. From about the age of 15, anytime I was not in the lineup I volunteered to “keep the book,” which is baseball jargon for documenting the outcome of each event (i.e. pitch, stolen base, error, etc.). While it was less exciting than playing, keeping the book enabled me to stay invested, and over time I developed a true fondness for delivering a neat and detailed final result – the box score.

In this case study I focus specifically on Major League Baseball (MLB) box scores as presented on ESPN.com. I made this decision because ESPN has emerged as a, possibly the, global leader in sports coverage.\(^1\) While a certain degree of domain-specific knowledge is required to interact with this organizing system, the primary organizing principles with which I am concerned are standardization, classification, authority and property persistence/essence. To my support my analysis, my artifact includes an encoding of Sports Markup Language as well as an exploration of robo-journalism.

**What is Being Organized?** In its most traditionally recognizable form, an MLB box score is an intentionally arranged 13x3 table that displays runs per inning as well as the total runs, hits and errors for the away team and home team, respectively. (If a game goes into extra innings, a new column is added for each additional inning.) By convention, the away team is always listed above the home team.

\[\begin{array}{ccccccccccc}
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & R & H & E \\
SF & 0 & 2 & 0 & 1 & 0 & 0 & 0 & 0 & 3 & 8 & 1 \\
KC & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 6 & 0 \\
\end{array}\]

**Screenshot 1: The tabular box score for Game 7 of the 2014 World Series**


Given that digital media outlets are not hampered by the spatial constraints often associated with print publications, in addition to runs, hit and errors, MLB box scores on ESPN.com include more granular information. Examples include hitting and pitching statistics by player, an English-language summary of each run scored and more general details about the game such as attendance and weather.

\(^1\) While it is certainly self-serving, it is worth noting that ESPN’s tagline is The Worldwide Leader in Sports.
Why is it being organized? ESPN.com organizes the information in order to present a summary of each MLB game. In that sense, the box score facilitates a set of interactions that enable users to parse through the data that determined the outcome of a given competition. Based on their preference, users can interact with the statistics and/or the English-language Scoring Summary.
**How much is it being organized?** All MLB box scores on ESPN.com adhere to a single design convention. The traditional 13x3 table appears at the top. Below that the screen is split into two columns. The away team’s statistics appear on the left, hitting before (i.e. above) pitching; the home team’s statistics maintain the same orientation except that they are displayed on the right. Under the hitting and pitching data the screen reverts to a single column where the *Scoring Summary* and *Game Information* are displayed, respectively.

Thus, the information is being organized according to well-defined structural relationships and highly-standardized encoded semantic relationships. In other words, the information is organized so that, once a user becomes familiar with the UI/UX layout, he/she can click on any MLB box score and intuitively know what they are seeing according to where it appears on the screen.

Screenshot 3: The red boxes and dotted line show the highly-structured, highly-standardized nature of the presentation

While the presentation layer is highly-standardized, the classification scheme is slightly more chaotic. For example, a single pitch can be classified as a strike and also a foul tip, but the order in which it is classified into a given category – strike or foul tip (or both) – does not matter because either way it is a strike. Though there are other simultaneously occurring categories – hit and homerun, for example – the point is not to explore each scenario. Instead, the point is to recognize that certain data points in the organizing system can be simultaneously classified in to multiple categories. Consequently, MLB box scores on ESPN.com do not seem to be organized according to either a faceted or hierarchical scheme.²

When is it Being Organized? During each MLB game, the box score on ESPN.com is updated in real-time. As a team scores a run, gets a hit, commits an error, etc. the appropriate data is added to the organizing system. In this context, the information is being organized at the time it is added to the organizing system (i.e. insertion).

However, there are times when a second phase of organization is required. These situations occur when a particular statistic already in the box score needs to be re-classified. For example, a pitcher’s performance can be re-classified from a win to a save.³ When this happens, the information is appropriately re-organized and then the box score is finalized (i.e. enters maintenance).

How or by Whom is it Being Organized? For the purpose of this case study, it is most appropriate to focus on authority in MLB. Though they operate behind the scenes, each MLB game has an official scorer,⁴ a “single authority [who exercises] effective control over the [game]….”⁵ This person is charged with ensuring that each statistic is classified according to the standards set forth by MLB.⁶ And even though many events are black and white – a homerun that goes over the wall, for example – there are others that are more nuanced. In these more complex cases – deciding between a hit and an error when an outfielder dives but misses the ball, for example –

² For more information about classification schemes see Chapter 7 of The Discipline of Organizing: Professional Edition edited by Robert J. Glushko (2014).
³ In Game 7 of the 2014 World Series Madison Bumgarner entered the game in the 4th inning as a relief pitcher for the San Francisco Giants. While he was initially credited with the win, upon further review, the official scorer determined Bumgarner should be credited with the save. The change was made the morning after the game and the box score was updated accordingly. Contextually, the difference between a win and a save is, to a certain degree, one of semantics (especially because the Giants won the game and the World Series). But, this example highlights that statistics in the box score can be re-classified. For more information see http://espn.go.com/mlb/playoffs/2014/story/_/id/11788539/world-series-scoring-change-gives-jeremy-affeldt-win-madison-bumgarner-save.
⁶ For specific information about the role and duties of MLB’s official scorers see http://mlb.mlb.com/mlb/official_info/official_rules/official_scorer_10.jsp.
the authority must try to remain as objective as possible. However, because the official scorer is a human, there is unquestionably a bit of subjectivity. Therefore, while it may seem as though the data in the organizing system is neutral, there is always, however slight it may be, the implicit bias of the official scorer.\footnote{For more information about biases in classification systems see Chapter 7 of The Discipline of Organizing: Professional Edition edited by Robert J. Glushko (2014).}

At a higher level of abstraction, the data in an MLB box score is being organized according to property persistence/essence.\footnote{For more information about resource properties see Chapter 4 of The Discipline of Organizing: Professional Edition edited by Robert J. Glushko (2014).} Simple data points – a hard-hit single into center field, for example – are organized by intrinsic static properties such that they are clearly perceivable and indisputable. On the other hand, more complex data points – for example, an outfielder who dives for a ball but misses it – are organized based on extrinsic static properties such that the result – hit or error – is assigned by the official scorer. In the complex cases, the higher the degree of “disputability” the more the data is organized according to extrinsic properties. To that point, the aforementioned example of a pitching performance being changed from a win to a save demonstrates the highest degree of “disputability” because the final outcome – save (not win) – was dependent on the authoritative yet subjective perspective of the official scorer.

\textit{Other Considerations.} In addition to the organizing principles already discussed, it is important to realize that, as it is experienced by a user, the information is organized by technology professionals at ESPN.com who maintain the websites on which box scores are displayed. Furthermore, from a capitalist perspective, ESPN.com organizes the data to earn advertising dollars, which they presumably do every time a user views a box score.\footnote{The white square in the top-right of Screenshot 2 and Screenshot 3 is where ads appear. The space has intentionally been made blank in order to avoid any/all potential copyright issues.}

\textit{Artifacts.} Created by the International Press Telecommunications Council (IPTC) in 2001, Sports Markup Language (SportsML) is “…an XML application for the exchange of sports data and statistics.”\footnote{http://www.xml.com/pub/r/1270 – Accessed on 15 Dec. 2014.} Although, or possibly because, each sport has domain-specific vocabulary (quarterback, pitcher, point guard, etc.), the metadata of SportsML is logically represented to describe many sports.\footnote{For information about the specifics of SportsML see http://www.iptc.org/cms/site/index.html?channel=CH0105.} Following is a basic encoding of the SportsML representation for Game 7 of the 2014 World Series (referenced above):

\begin{verbatim}
7 For more information about biases in classification systems see Chapter 7 of The Discipline of Organizing: Professional Edition edited by Robert J. Glushko (2014).
8 For more information about resource properties see Chapter 4 of The Discipline of Organizing: Professional Edition edited by Robert J. Glushko (2014).
9 The white square in the top-right of Screenshot 2 and Screenshot 3 is where ads appear. The space has intentionally been made blank in order to avoid any/all potential copyright issues.
\end{verbatim}
After the event has been encoded in SportsML, the result can be rendered in HTML for a clean, tabular display:\textsuperscript{13}

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>San Francisco Giants</td>
<td>3</td>
</tr>
<tr>
<td>Kansas City Royals</td>
<td>2</td>
</tr>
</tbody>
</table>

Beyond the UI/UX presentation, because it is highly-structured, SportsML can also be used to automatically generate recaps of a given contest. In other words, SportsML facilitates robo-journalism.\textsuperscript{14} Referencing the encoding above, it is easy to see how a piece of journalism software could scan the metadata and automatically create a headline that reads Giants Narrowly Beat Royals. But, in order to auto-generate a complete box score, even in the tabular format, the metadata would necessarily have to include other required attributes such as innings, hits, errors, etc. Furthermore, in order for the metadata to be utilized to auto-write a full recap quickly and accurately, the corpus of the software would need to have enough English-language words to describe the ebbs and


\textsuperscript{13} Ibid.

flows of the game. The syntax would most likely have to be domain-specific so that the auto-scribed story makes sense – for example, a baseball player who slam dunks a goal would be quite an anomaly – but that could be handled by building a corpus for each sport.

Interestingly, it seems that ESPN.com has not implemented SportsML for MLB box scores. For example, neither “sports-event” nor “event-outcome” – two of the SportsML elements in the encoding above – can be found in the page source for the box score for Game 7 of the 2014 World Series on ESPN.com. While it would be an involved exercise, if ESPN.com were to adopt SportsML it would be necessary to compile crosswalks, “…which are equivalence tables that relate resource description elements, semantics, and writing systems from one organizing system to those of another.” Though this is pure speculation, it is doubtful that ESPN.com chooses not to employ SportsML because the crosswalk exercise would be difficult and time-consuming.

Instead, it is more likely that ESPN.com elects to maintain proprietary code (as opposed to using SportsML, which is open source,) for two reasons. The first is so that the company can continue to employ sports writers who can promote the biases of the corporation. Humans, it seems, are quite good at towing the company line. (When they do not, the consequences are very real.) The second is so that other media outlets cannot scrape the data that ESPN.com took the time, and spent the money, to organize. If this was the case, and other sites presented the information as their own, there would be no compelling reason for users to view a box score on ESPN.com. And, as mentioned in Other Considerations, since advertisements are prominently placed on each page that displays an MLB box score, if people stopped trafficking ESPN.com the company would stand to lose revenue. Therefore, while it may seem counterintuitive, such that it goes against the idea of adopting universally-applicable standards, ESPN.com will most likely avoid implementing SportsML so that it can maintain its self-appointed status as The Worldwide Leader in Sports.

15 For more information about building a corpus, see “Techniques and Applications for Sentiment Analysis” by Ronen Feldman (2013).
16 The page source for the website (http://scores.espn.go.com/mlb(boxscore?id=341029107) was examined in Mozilla Firefox by clicking Tools → Web Developer → Page Source.