

**Original File A and File B Profiles (Part 1, Step 1 )**

1. Well-formed XML File A contains 214 lines, 1211 words, and, what seems to be, an unordered list of 8 customer complaints (not ordered either by complaint ID or date). It uses XML version 1.0 with the “Unicode Transformation Format – 8-bit” (UTF-8) character encoding. Each nested element starts on a new line with proper indentation. Therefore, the file has a lot more lines than File B. The order of nested children of element “complaint” is strictly the same for the first 7 complaints (provided that some of them have 0 or 1 occurrence), but it is significantly “reshuffled” for the 8<sup>th</sup> complaint. MD5 checksum: abc4665edea350247bdcacb90dfb9b04
2. Well-formed XML File B contains 116 lines, 1210 words, and the same list of 8 customer complaints (ordered per File A). As File A, it also uses XML version 1.0 with the UTF-8 character encoding. The file format is different from File A: some third-level nested elements are “glued” together without new lines or indentation (inside element “complaint”). This makes the information look more compact, but you have to be more careful when you read the file visually. MD5 checksum: 1ede428d670ef18e68a3c326efb91bc4

Below is a summary of syntactical differences between File A and File B:

- a) File A element “submitted” with attribute “via” became an additional attribute “submissionType” of File B element “complaint”. The purpose was, probably, a more compact and clear data representation. Apparently, this was done for element “complaint” id=2364257 in a wrong way, where a dangling, empty, attributeless, nested “submitted” element was left inside element “complaint”, and attribute “submissionType” was not added to parent element “complaint” although the information is present in File A (information loss). The same kind of information loss occurred in File B complaint id=837784 except for the fact that there is no dangling element “submitted” left behind;
- b) Attribute “timely” of element “response” started to have values “yes” or “no” in File B in some cases (preserving Y/N in others), while it was always Y/N in File A in all the cases. And in any case, its value should be harmonized with a similar type of value for attribute “consumerDisputed” which is Y/N in File B;
- c) Attribute “timely” of element “response” is missing in File B complaints id=837784 and id=14038 although this information is present in File A (two counts of information loss);
- d) There is one comment about complaint id = 837784 saying that a specific employee changed the record (supporting provenance). This is useful and should probably be made a separate child element “comments” nested under parent element “complaint”;

- e) The order of attributes “type” and “date” in File B element “event” varies while it is always the same in File A. This should not be an issue for DTD validation purposes, but may be a point to consider to beautify File B.