Andrew Nedilko
Final Project
IS 531: Foundations of Data Curation
Fall 2017

Answers to Reflection Questions (Part 1, Step 5)

a) In order to canonicalize the two documents, first I had to make decisions for each discrepancy between the files noted in Step 1, Part 1 (see a separate file "FileA_and_FileB_Profiles_Part1_Step1.pdf"). I made the following decisions:

- For canonicalization purposes, each element, start tag, end tag has its own line; no line breaks in elements of type PCDATA. I also had to normalize the text in the consumerNarrative elements as it contained some extra symbols in one file vs. the other (extra semicolons, two single quotation marks instead of one double quotation mark, etc.);
- File A and B treat the complaint submission type differently: it is an element in File A and an attribute in File B. In order to canonicalize the two files and test the idea that they represent the same data, I had to do something about it. Otherwise, the character sequences and checksums would be different by default. Since File B is the new system, I added attribute "submissionType" to element "complaint" and removed a separate element "submitted" in File A. In File B, I added the missing attribute "submissionType" to elements "complaint" id=2364257 and id=837784 and removed the "submitted" child element in elements "complaint" id=2364257 to eliminate information loss;
- The values of attributes "timely" and "consumerDisputed" are only "Y" or "N" (no "yes" or "no");
- I added missing attribute "timely" to element "response" in File B complaints id=837784 and id=14038 from File A to avoid information loss;
- Per instructions, I removed the comment in File B for element complaint id = 837784 which seems to support provenance. Although, I still believe it should be made a separate child element "comments" nested under parent element "complaint";
- Attributes "type" and "date" in File A and File B element "event", as well as all other attribute/value pairs, were listed in the alphabetical order.

I also took additional actions as described in the XML canonicalization example (video 4, week 9). They included the following:

- Although the header says that both files already use the UTF-8 single character encoding, I resaved the files as UTF-8, just to be on the safe side. While doing this, I also normalized line ends (Windows style) and end tags (to be similar between the two files).
- I removed all comments, tabs, non-significant spaces (quite a few of those, especially in File B compared to File A, even inside the quotation marks), so that both files have the same format.
- I propagated all attribute defaults from DTD to the elements themselves. In doing so, I had to normalize attribute values and decide whether they should have default values or just have a generic type (based on the discrepancies noted in step 1, part 1, as noted above).
- I removed the minimal internal DTD from File B. Because of this, &redaction had to go back to XXXX. Couldn't find any declarations except for the first line of each file, but I left it as it is meaningful.

The checksums after the canonicalization are "397655e7f66948d272b45777453747d6" for both files. Which means that the two files, and the two datasets contained in them, are identical.

b) How does the way data is represented impact reproducibility?

As we learned from the lectures, reproducibility intends to document not only data collection and management, but also processing and analysis. The ability to reproduce the results is a required condition for the progress. This is the way the result of a study can be validated or rejected. Data representation plays a crucial role in this process because if the same data is represented in different ways, this can make processing and analysis much more complicated and time-intensive. On the other hand, if two datasets seem to be similar, but are somehow different, the results of studies based on them can be very misleading and unjustified.

c) How may your canonicalization support the overarching goals of data curation (revisit objectives and activities of Week 1)?

Data curation deals with all aspects of data management to provide an efficient and reliable support for the analysis and reuse of data. My canonicalization confirmed the identity of the two datasets which makes data analysis based on them valid and substantiated. Analysis is more precise and justified if the data you are working with are more clear and understandable.

I believe that my canonicalization directly supports the following areas of curatorial activity:

- Preservation by ensuring that data will be understandable and usable in the future and by establishing the identity of the two datasets;
- Discoverability by supporting the ability to search for and locate relevant data;
- Organization through the use of an adequate data model/schema and adequate standards including syntax and semantics standards, employment of abstraction and indirection to manage data;
- Access by ensuring the ability to retrieve and distribute data;
- Workflow by ensuring the ability to systematize work with data;
- Identification by ensuring the ability to identify data through an identifier system;
- Reformatting because a well-formed XML document with a valid DTD ensures the reformatting for use by different tools or to match new format standards;
- Sharing by supporting easy distribution of data between researchers, teams, and institutions overcoming common obstacles on this way;
- Communication because a valid document enables representation, publishing, and visualizations that provide insight, and
- Reproducibility because the process of canonicalization confirmed the identity of the two datasets which will have a positive impact on the reproducibility enabling the respective data processing and analysis with subsequent validation of results.

d) Which additional curation activities would you recommend to enhance the data set for future discovery and use?

I would recommend modification and security: ensuring proper access to data and tracking all the changes is very important in the process of data management. Also, organizing proper storage of data is a critical factor in this process.