Connor Segneri – segneri3

UIUC – Foundations of Data Curation

10/15/19

Final Project Writeup

Part 1

File Profiles

File A (Old System)

The format of File A is in xml. It contains a list of consumer complaints. Each complaint has an id, event, product, issue, company, submitted by value, and a response. It also has an optional consumer narrative. Some of these elements are made up of even further elements, and some elements have various attributes. “product” is made up of two other elements, “productType” and “subproduct”. “issue” is made up of “issueType”. “company” is made up of “companyName”, “companyState”, and “companyZip”. “response” is made up of “responseType”. At the top of the file, the only things specified are the xml version, which is “1.0”, and the encoding, which is “UTF-8”.

MD5 Checksum

- D30CBA6B00308A87FA3A384799C5FAF7

File B (New System)

File B, much like File A, contains an xml list of “consumerComplaints”. The header of this file contains a bit more information. The header contains a doctype with a list of a single entity, “redaction”. Each complaint contained in the file contains elements and attributes that are like File A, but with several changes. “submissionType” is now an attribute for the “complaint” element, and not an element on its own. Also, the spacing is different between the elements. There are also various references throughout the consumer narratives to the “redaction” element defined in the doctype. The string assigned to “redaction” is “XXXX”.

MD5 Checksum

- 47677272E76E1F4332AFE859347C8695

Checksum Comparison

Looking at the two checksums above, we can see that they are not the same, which means we can not yet say that the two data sets have the same data representation.

DTDs for Both Files

The DTD for File A is in the “ OldSystem_WithDTD.xml” file, and the DTD for File B is in the “NewSystem_WithDTD.xml” file.

XML Canonicalization

The canonicalized files are in “OldSystem_Can.xml” and “NewSystem_Can.xml”.

Final DTD

The Final DTD is in the file “FinalSystem_WithDTD.xml”.

XML File Validation

All xml files validate according to http://xmlvalidator.new-studio.org/.

Reflection Questions

a) Describe your process for canonicalization (i.e., decisions, actions, representation selection, attribute issues, provenance decisions). Report the checksum values after canonicalization.

The differences between the files start with “submissionType”. In the old system it was an element, in the new system it’s an attribute of “complaint”. The whitespace between systems was also different in various ways, so I normalized that to each element being separated by a single new line. If a product contained other elements, then the opening and closing element tags were separated from the internal elements by a new line. The “timely” attribute uses “Y” and “N” instead of “yes” and “no”, this is how the old system does it, and it’s like other attributes in the files. I also made sure there was no trailing whitespace in the element values. During this process, I replaced “&redaction” with its actual value, “XXXX”.

The checksum for the old system’s canonicalization is:

38627028a8ff07e3ab43c0c9aceaf67a

The checksum for the new system’s canonicalization is:

38627028a8ff07e3ab43c0c9aceaf67a

The checksums are equivalent, and it seems as though the representation syntax of the old system and the new system are the same.

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

Reproducibility supports the ability to reproduce results, ensuring scientific validity and reliability. If the way that data is represented changes, then there might not be a way to get the same report/analysis/conclusions from the data that were previously achieved. If there were, for example, a report that showed how many complaints were made to a certain company, and the underlying representation of the data was changed, that report could give incorrect information, or fail to render entirely. Reproducing results like these means documenting every bit of processing and analysis. This way, when representations change, the solutions to reproducing past insights is clear.

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

Some of the overarching goals of data curation are Preservation, Access, Sharing, and Communication. Preservation is ensuring that data will be understandable and useable in the future. Canonicalization showed that the old system and the new system have the same representation syntax. We know that the data representation was preserved in the transition from the old system to the new, something we couldn’t assume. Access is the ability to retrieve and distribute the data. Now that we know the new system maintained the data representation from the old system, we can give access to the new system to other employees that we couldn’t before, now that we know they’ll be able to access everything they could in the old system. Sharing is like access, and it specifically supports sharing data between researchers, teams, and institutions. Canonicalization normalized our data representation, so when we share our data sets with others, they can expect the same xml format each time. Communication is the support of representation, publishing, and visualizations that provide insight. Canonicalization helped us to prove that any past

visualizations and publishing's can still be reproduced using the new system, since the representation of data is the same.

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

I would suggest writing more documentation as to why the company is switching systems. I would also recommend some prior analysis of the old system, and further documentation into its shortcomings. Some policy documentation would also be good concerning objectives, procedures, practices, and data formats.

Part 2

Data Curation Memo

Data curation is a necessary part of data collection, data analysis, and data science. In this memo, I'll introduce the idea of data curation, and explain why I believe the previous statement is true. Data science is concerned with all aspects of the creation, management, analysis, and communication of data focusing particularly on the application of computational methods to digital data. Data curation is essentially the aspects of data science that are not focused on data analytics. Data curation is the active and ongoing management of data through its lifecycle of interest and usefulness to scholarship, science, and education. There are many curatorial actions that I believe will save the company time, ensure the accuracy of our data, and the accuracy of any reports/analyses/conclusions that stem from our data. The first aspect I'd like to address is Provenance.

Provenance means identifying what inputs, processes, and calculations are responsible for data values. Any sort of data pipeline relies on provenance for engineers to be able to understand how certain data values are computed. If there isn't extensive documentation or descriptive metadata concerning the transformation of data into new data, multiple issues arise. One such issue is that any engineers who aren't familiar with the data pipeline rely on other engineers to explain the process to them. This assumes that there will always be engineers who remember the process, and remember it perfectly, which is never a safe bet. Another reason why provenance is so important is for troubleshooting. It could be that earlier in a data pipeline a data set was significantly changed, breaking a subsequent transformation that creates a new data set. Without documenting which input data sets are used to make up the now corrupted data set, engineers waste time trying to figure out the problem without having easy access to a resource that tells them where they should start their search.

One way to go about implementing the idea of provenance is through policy. Policy means to specify objectives, procedures, practices, and formats. Policy will range from things having to do with regulation, law, property, data formats, metadata, and more. One of the biggest issues being legal policy. If we don't follow legal policy, and one of our engineers makes a mistake that slips through to production, that could be the end of our company. Creating, identifying, and following policy is import so that our engineers know how to document properly, use proper data formats, and generate correctly formatted metadata.

Speaking of metadata, having a documented methodology for generating metadata is more important that some might think. Metadata is used to make sure that data is usable and understandable. Properly implemented metadata supports searching the data, getting the data into the right format, and identifying what data certain users have access to. Without proper metadata policies, engineers either won't utilize the benefits of metadata, or everyone will have a different idea about how to generate and utilize metadata, which causes unnecessary chaos.

Preservation is the last idea I want to touch on. Preservation ensures that data will be understandable and useable in the future. This is the main idea behind a lot of data curatorial activities. If we don't spend time working on this now, then the future state of our engineers and the company will suffer. Documentation, metadata, and policies all act to maintain the quality, reliability, and accuracy of the data that we process. I know it means working

more and spending more time now, but I believe data curatorial work is necessary for us to have any faith in the insights we pull from our data.

Resources

https://tools.chilkat.io/XmlCanonicalize.cshtml#xmlFormat

http://xmlvalidator.new-studio.org/

http://onlinemd5.com/

https://xml.mherman.org/