Connor Segneri – segneri3

UIUC – Foundations of Data Curation

10/15/19

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 a number of 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 in this file than the previous. 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 the files are not the same.