The process:

As we saw previously, the checksums of the two given files A and B did not match when used in their original form. Thus the process of canonicalization will be carried out now after observing the mismatch. During this process the following changes were made to each of the files:

- Everything before the root element's opening tag was removed which included the internal DTD and the XML declaration
- Whitespaces outside the document's root element was normalized
- Comments that were present within the documents were removed
- The document was already encoded in the UTF-8 encoding, so nothing was done on that end
- Empty elements were converted into start-end tags. For example: In file A: <submitted via="Referral"/> was converted into <submitted via="Referral"></submitted>
- Normalization of whitespace has been done in start and end tags. All whitespaces have been retained as is between consecutive start tags and between consecutive end tags.
- All the attributes were normalized so any white space inside a tag between attributes is replaced by a single space and all attribute values are surrounded by double quotes (") i.e no line breaks and exactly one space between two attributes.
- Any space between the double quotes in an attribute value was removed. Example type="sentToCompany "(to be read as ="sentToCompany(space) ") was changed to type="sentToCompany".
- A Lexicographic ordering was imposed on the attributes of each element. For example <event type="received" date="2014-03-12"></event> was changed to <event date="2014-03-12" type="received"></event> because attribute "date" comes first alphabetically before "type".
- Encoding of special characters was done as character references in text. For example M&T changed to M&T

Even after canonicalizing the two files, there appears to be a difference in the checksums. After canonicalization the checksum values are as follows

1) File A was "fbee357ec61f7bf1867e41c3e03327cc"
2) File B was "b2fa8d550f7962c1f1dc8d55b83976ab"

They do not match. This shows us that there is a necessity to further check the formatting between the two files and identify dissimilarities. Hence the two files were closely scrutinized to find minor differences between them with respect to "**structure/formatting**" of the xml documents. Some of them were :

1) The SubmissionType is used as an attribute for the element complaint in file B whereas it is present as an element in file A.
2) The two attributes "consumerDisputed" and "timely" within the response element have values assigned to them corrsesponding to a yes or a no. But the values used do not

appear to be consistent as both “Y” and “Yes” have been used to mean yes and “N” and “No” have been used to mean no..

The decision : Since there are minor differences between the two files with respect to formatting, there is a need to consider one format and standardize the files according to the format. Since **File A was the original file** that was used by the company, there will be no issues relating to missing data and thus I will be considering it as the benchmark for the rest of the project. One other reason file A can be better than B is it considers submissionType as a child element than as an attribute and this makes the document more readable, easy to maintain and make changes. Hence file A’s DTD will form the basis for creating file B’s changes.

The changes I will make now to the initial canonicalized files are:

1) The only change made to file A will be to standardize the value that the attributes “consumerDisputed” and “timely” take on. The Y and N are changed to Yes and No respectively to maintain uniformity within and across the files. This file will then be considered as the **final file** with which we will try to compare file B with. The same formatting will be done in file B as well.
2) All the “SubmissionType” attributes are converted into child element and named as “submitted” in file B (The name is changed to keep it consistent with naming convention in file A)

At this point I ran the checksums again to see that they still do not match.

1) File A read “aa590f8fd5a6591cc6485d4cbfce05a5”
2) File B read “60cc78148d81a372a4bbe885e2e0f0d1”

(Note: This intermediate file is attached in the final project submission by the name “File B intermediate”)

This actually proves that the data transfer that the company did was “**not completely successful”** in moving the data from the old system to the new system.

Now it is seen that this happens due to “**data loss**” while migrating. File B suffers loss of data in the following places:

1) In the complaints with id attribute numbers 14038 and 837784 (i.e the 7th and 8th in the file) the attribute “timely” within the response element is missing.
2) In the complaint element with id 2364257 there is an empty child element called “submitted” which does not contain information about the type of submission. The submissionType is not mentioned as an attribute within the complaint element either(which is true for other cases)
3) Similar mistake is also found in complaint with the id attribute 837784 where the submissionType is not recorded.

Thus it is important for the company to note that the data from the old system will be successfully migrated to the new system if they add these missing data and make the necessary corrections.

I hence added the following two things to file B; the things that were present in file A but were missing in file B:

1) Attribute "timely" was added to the "response" child element for complaint ids 14038 and 837784
2) The empty "submitted" element within complaint id 2364257 was assigned appropriate value as seen in file A
3) A "submitted" child element is added to the complaint with id 837784 and its corresponding value from file A has been assigned.

After making the above changes to file B, the checksums of the two files match perfectly. They are shown below:

1) File A read "aa590f8fd5a6591cc6485d4cbfce05a5"
2) File B read "aa590f8fd5a6591cc6485d4cbfce05a5"

(Note: These files are attached in the final project submission by the names "File A final" and "File B final")

Question 5b) How does the way data is represented impact reproducibility?

The DTD for the final canonicalized file can be used as a standard by the company even when the system has to undergo changes in the file system in the future. Although the migration of data has to strictly adhere to the DTD format to control information loss, there is still scope for improvement of schema; like carefully converting the elements to attributes or vice versa, without having to suffer significant data loss.
Further, to add to this, the entire process of canonicalization is well documented and analyzed to make decisions. This approach can be used by other teams who are undergoing a similar change in the company as well.

Question 5c) How may your canonicalization support the overarching goals of data curation?

1) Organization: It uses an appropriate data model and DTD schema that is relevant for the source documents and also uses abstraction and indirection to manage data.
2) Preservation: The process has all the required documentation regarding the semantics of the attributes and the elements in explained detail. It is thus clearly understandable and reusable in the future.
3) Discoverability: The above schema allows the data to be easily scanned through/spotted because of its hierarchical arrangement.
4) Access: On retrieving the metadata, it is very easy to understand the format of the document.

5) Workflow: Each step in the process of canonicalization is well explained and documented. With the help of the documentation, the canonicalization process can be well executable.
6) Identification: The similarities and dissimilarities of both files A and B are discussed. The DTD schema developed for the files is both authenticated and validated(All the schema constraints, syntax and schematics are met)
7) Modification: If the data has to be changed, the documentation provides a good explanation of how and where the changes can be made.

Question 5d) Which additional curation activities would you recommend to enhance the data set for future discovery and use?

I would recommend the “Integration” activity in the future if there is a need to make use of the formatting used in the new file system(file B). I.e if for some reason the company decides that having “SubmissionType” as an attribute makes more sense than having it as an element, then we will need to integrate files A and B(because file B cannot be used as is due to missing data during migration.) For this there needs to be a schema alignment between files A and B as well prior to integration. This leads us to the task of “Reformatting”, i.e using the new file (file B) format to store data. Not just the present new format, but also adapting and changing to any new tools or standards in future.