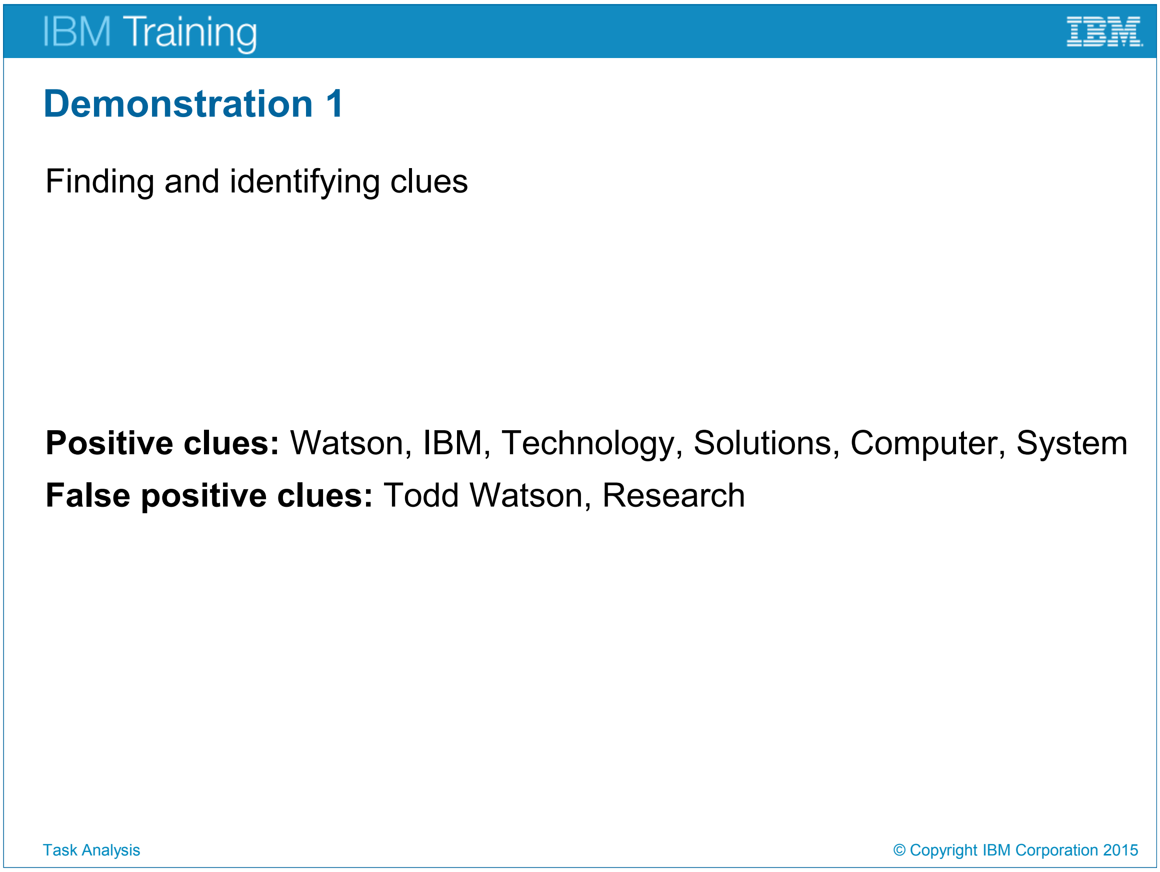


IBM Training

Demonstration 1

Finding and identifying clues

Positive clues: Watson, IBM, Technology, Solutions, Computer, System

False positive clues: Todd Watson, Research

Task Analysis

Demonstration 1: Finding and identifying clues

Demonstration 1: Finding and identifying clues

Purpose:
This demonstration will show you how to find and identify clues that are needed for the extractor. In real life, this process would typically be done with assistance from a subject matter expert, or someone who is familiar with the documents that you are examining. Prior to starting this demonstration, ensure that all the necessary Ambari services are up. If you had just completed Demonstration 1, you are in good shape. Otherwise, refer to demonstration 1 to get that set up.

User ids / Passwords

OS:	**biadmin/biadmin**
Root:	**root/dalvm3**
Ambari:	**admin/admin**
BigInsights Home:	**guest/guest-password**

Ambari Services Required:

- HDFS
- MapReduce2
- YARN
- Knox (also start the Demo LDAP service)
- BigInsights - Text Analytics
- BigInsights - Home

Task 1. Finding your way around the Web UI.

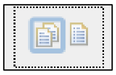
1. With the required services started, open up a new browser (or a new tab).
2. Go to the BigInsights - Home page. Use the bookmark saved in the Firefox browser, or this URL:

   https://ibmclass.localdomain:8443/gateway/default/BigInsightsWeb/index.html#/welcome
3. Click on Text Analytics to load up the Web UI.
4. You have used this in the first demo, but let's spend a little more time on the Web UI to make sure you know your way around. If you feel comfortable enough, you may skip this task. The left side of the UI has your **Projects** and **Extractors**. Click on the **Ed Demo** project to load it (if it wasn't already loaded).

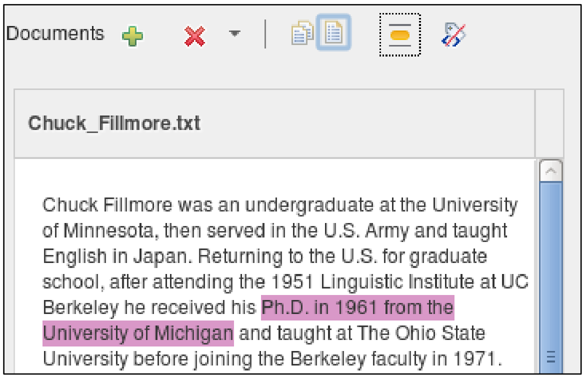
   This loads all your extractors onto the canvas. It also loads the documents that were used in that project.

5. Click on the **Extractor** tab to see the list of the pre-built and custom-built extractors. You can drag and drop these directly onto the canvas to start using them.
6. On the canvas, select the **Degree** extractor.
7. Expand the **Extractor Properties** pane to see its settings. You may need to resize by click and dragging the pane. Play around with this to get comfortable in resizing the panes.

   Note: You can only resize if the panel is expanded.
8. Under the Extractor Properties, there are three sub-tabs: **General**, **Settings**, and **Output.** Click the **General** tab (if it isn't already on it).
9. On the **General** tab, you can edit the name, provide a description, or define some tags to assist in being more easily searchable among the Extractor catalogs. We will not do anything here, this is just for your information.
10. Click on the **Settings** tab. On here, you can modify the terms in the dictionary (in this case) or if it was a different extractor, modify the settings of that one.
11. Click on the **Output** tab. Here is where you can specify the columns from the extractor.
12. On the canvas, click on the **Education History** extractor and run it.
13. Go ahead and collapse the **Extractor Properties** and expand and resize the **Results** pane so that it is more visible.
14. Each tab on the results pane comes from a single extractor. In our case, we have a single union of multiple extractors, so we have single tab. Within that one tab, however, we have multiple results, one for each of the extractors that made up that union. Examine the results to see the various columns.
15. Click on any row and you will see that the results are highlighted within the document on the **Documents** pane (on the right).
16. Remember, you have the option to export your results as a CSV file for further analysis with a different tool.
17. On the **Documents** pane, you can toggle between single document view and multiple document views. Go ahead and click on it to see it in action.

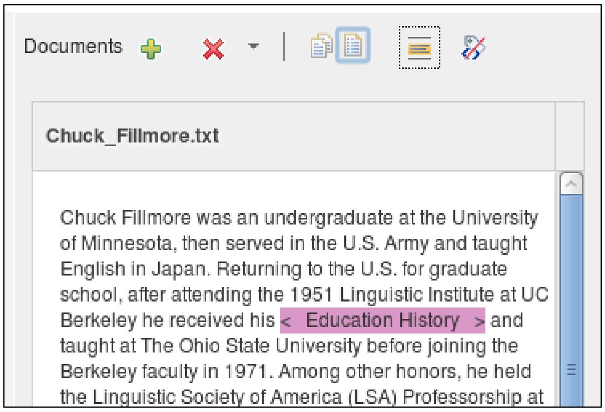
18. Next to that is another button, **Show Extractor Name.** This is a nice little feature that tells you which extractor found the results. For example, select one of the rows from the Results pane.



Chuck_Fillmore.txt

Chuck Fillmore was an undergraduate at the University of Minnesota, then served in the U.S. Army and taught English in Japan. Returning to the U.S. for graduate school, after attending the 1951 Linguistic Institute at UC Berkeley he received his Ph.D. in 1961 from the University of Michigan and taught at The Ohio State University before joining the Berkeley faculty in 1971.

19. Now click the **Show Extractor Name** button to see which extractor it was:



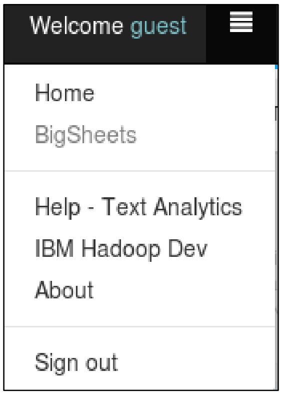
Chuck_Fillmore.txt

Chuck Fillmore was an undergraduate at the University of Minnesota, then served in the U.S. Army and taught English in Japan. Returning to the U.S. for graduate school, after attending the 1951 Linguistic Institute at UC Berkeley he received his < Education History > and taught at The Ohio State University before joining the Berkeley faculty in 1971. Among other honors, he held the Linguistic Society of America (LSA) Professorship at

Obviously, in this case, we only had one extractor, but if you ran with multiple extractors, you can use this to find out which one captured that result. This can help with debugging if you end up finding terms that should or shouldn't be part of the result set.

20. Finally, the third button is the **Remove tag / Remap tag.** This is used for documents where you may have tags, such as XML documents.

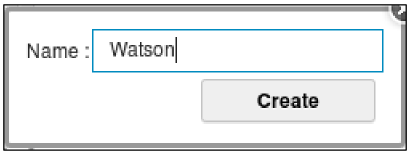
21. If you need additional help, at the upper right corner, there is a dropdown icon. Click on that and you can visit the help section for Text Analytics.



This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Task 2. Creating the Watson project.

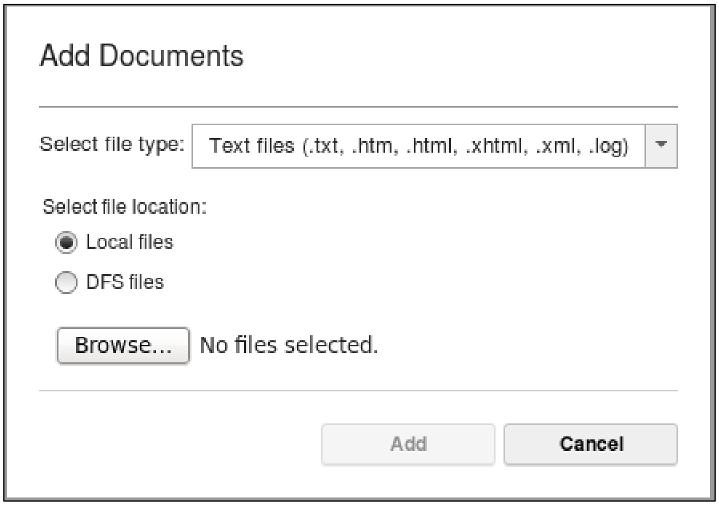
1. On the **Project** pane, click the **green plus**.
2. Specify the name Watson for the project.



Name : Watson

Task 3. Loading the data files.

1. On the **Documents** pane, click the **green plus**.



Add Documents

Select file type: Text files (.txt, .htm, .html, .xhtml, .xml, .log)

Select file location:

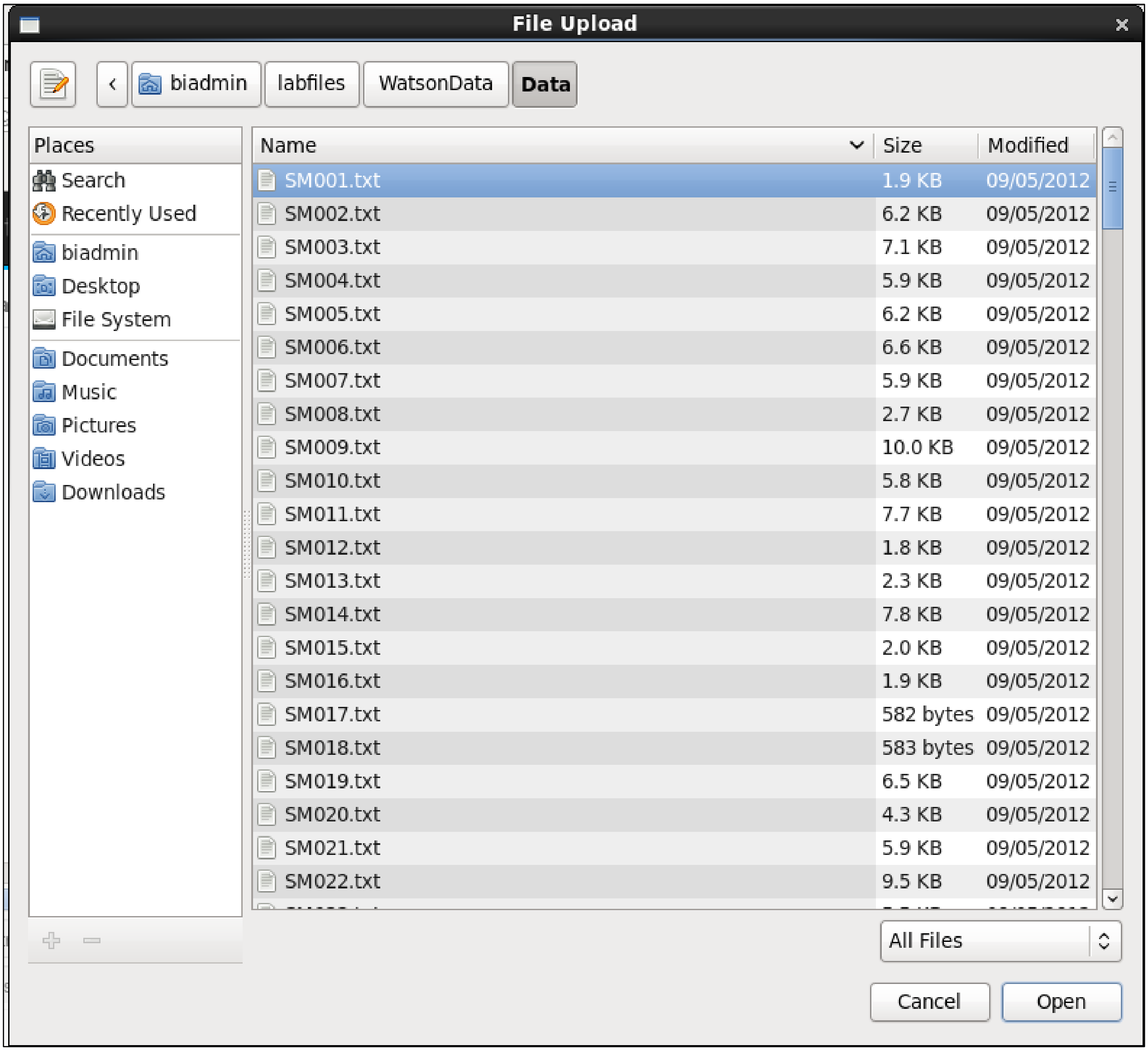
● Local files

○ DFS files

Browse... No files selected.

2. Specify the file type as **Text files** and the file location as **Local files**. Click **Browse** to select the files.

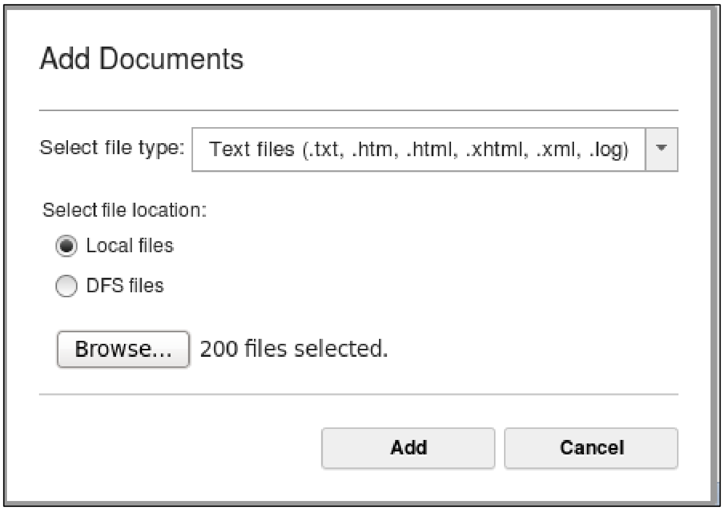
3. Navigate to **/home/biadmin/labfiles/WatsonData/Data/**.



Name	Size	Modified
SM001.txt	1.9 KB	09/05/2012
SM002.txt	6.2 KB	09/05/2012
SM003.txt	7.1 KB	09/05/2012
SM004.txt	5.9 KB	09/05/2012
SM005.txt	6.2 KB	09/05/2012
SM006.txt	6.6 KB	09/05/2012
SM007.txt	5.9 KB	09/05/2012
SM008.txt	2.7 KB	09/05/2012
SM009.txt	10.0 KB	09/05/2012
SM010.txt	5.8 KB	09/05/2012
SM011.txt	7.7 KB	09/05/2012
SM012.txt	1.8 KB	09/05/2012
SM013.txt	2.3 KB	09/05/2012
SM014.txt	7.8 KB	09/05/2012
SM015.txt	2.0 KB	09/05/2012
SM016.txt	1.9 KB	09/05/2012
SM017.txt	582 bytes	09/05/2012
SM018.txt	583 bytes	09/05/2012
SM019.txt	6.5 KB	09/05/2012
SM020.txt	4.3 KB	09/05/2012
SM021.txt	5.9 KB	09/05/2012
SM022.txt	9.5 KB	09/05/2012

4. Select all the files. Use **CTRL + A** to select all the files and click **Open**.

5. Click **Add** to add the files.



Add Documents

Select file type: Text files (.txt, .htm, .html, .xhtml, .xml, .log)

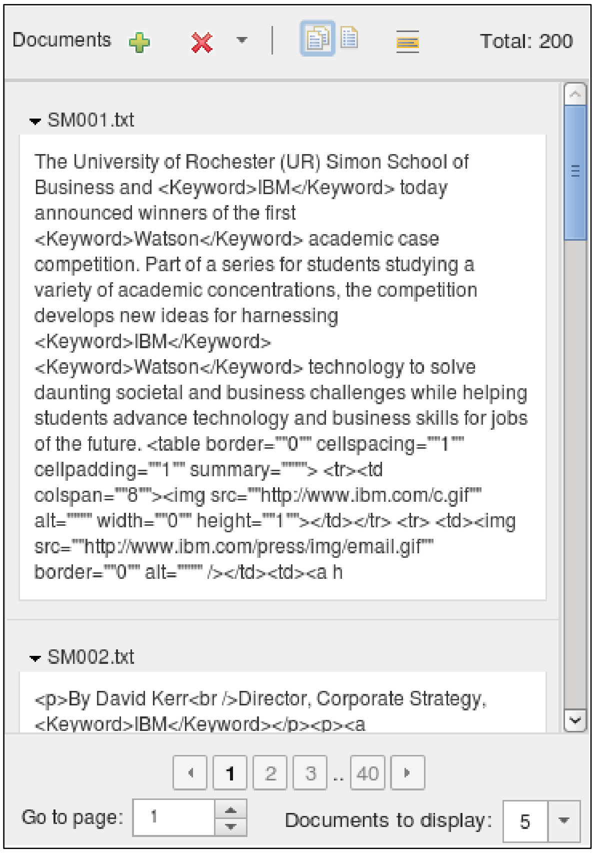
Select file location:

● Local files

○ DFS files

Browse... 200 files selected.

6. The documents are loaded.



Documents Total: 200

SM001.txt

```
The University of Rochester (UR) Simon School of
Business and <Keyword>IBM</Keyword> today
announced winners of the first
<Keyword>Watson</Keyword> academic case
competition. Part of a series for students studying a
variety of academic concentrations, the competition
develops new ideas for harnessing
<Keyword>IBM</Keyword>
<Keyword>Watson</Keyword> technology to solve
daunting societal and business challenges while helping
students advance technology and business skills for jobs
of the future. <table border=""0"" cellspacing=""1""
cellpadding=""1"" summary=""""> <tr><td
colspan=""8""><img src=""http://www.ibm.com/c.gif""
alt="""" width=""0"" height=""1""></td></tr> <tr> <td><img
src=""http://www.ibm.com/press/img/email.gif""
border=""0"" alt="""" /></td><td><a h
```

SM002.txt

```
<p>By David Kerr<br />Director, Corporate Strategy,
<Keyword>IBM</Keyword></p><p><a
```

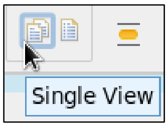
Go to page: 1

Documents to display: 5

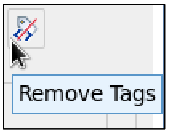
Task 4. Identifying and creating a list of the clues.

In this task, you will be creating a list of clues that you will use to create your extractor. Your test data consists of a number of files that are actually a collection of blogs and news posts retrieved from various social media sites using the BigInsights sample Boardreader application. Each post is stored in an XML encoded format. You use this test data to find examples of the type of information that you want to extract and build your extractors based on those examples.

1. Locate the file **SM001.txt**. Select that file and choose the **Single View** to show only one document at a time.



2. Next, make it easier to read by removing the tags by clicking on the **Remove tag** icon.



3. This is a copy of the text:

   The University of Rochester (UR) Simon School of Business and IBM today announced winners of the first Watson academic case competition. Part of a series for students studying a variety of academic concentrations, the competition develops new ideas for harnessing IBM Watson technology to solve daunting societal and business challenges while helping students advance technology and business skills for jobs of the future.

4. Since the goal of the task is to find social data that references IBM Watson, the first snippet of interest would naturally be the word Watson. Make a note of this word in a Notepad or a text editor of your choice. We'll keep a running note here:

   Positive clues: Watson

5. It is easy for you, as a human being, to scan through these files and find those that are referencing the Watson technology as opposed to someone’s name or a place. But that same innate capability does not exist for a computer. You are going to have to give the computer both positive and negative clues for it to be able to recognize the appropriate Watson reference.

The first reference to Watson in the text was related to a competition. The second reference was IBM Watson technology. This is a reference in which we have an interest. And there are two clues that are of value, IBM and technology. It is the word Watson in context with these clue words that allow us to make the assumption as to the meaning of the word, Watson, used here.

Positive clues: Watson, IBM, Technology

6. Locate the SM010.txt file.
7. Examine the file and take note of the words *Solutions* and *computer.* These clues also relates to the Watson technology and will help the computer figure out if the Watson within the document is the Watson we want.

   Positive clues: Watson, IBM, Technology, Solutions, Computer
8. Locate the SM005.txt file.
9. Examine this file and take note of the word System.

   Positive clues: Watson, IBM, Technology, Solutions, Computer, System
10. Locate the SM011.txt file.
11. Examine the document and take note of the word Jeopardy

    Positive clues: Watson, IBM, Technology, Solutions, Computer, System, Jeopardy
12. Locate the SM063.txt.
13. Here we will look for some negative clues, or clues that may give false positives (e.g. returning Watson where it does not have anything to do with technology, but rather, a person's name or something of that nature).

    False positive clues: Todd Watson
14. Locate the SM121.txt. It's on page 25 if you are searching by page number.
15. In this file you have Watson Research Center in Yorktown Heights. Research would be another good false positive:

    False positive clues: Todd Watson, Research
16. At this point, we have enough information to work with to demonstrate the capability of BigInsights Text Analytics.

Results:
At the end of this demo, you should be able to identify clues that are needed for the extractors. You understand that typically, this process would involve someone who is familiar with the documents, such as a subject matter expert.