VOLUME 102 NUMBER 3 FALL/WINTER 2018
JUDICATURE
Published by the Duke Law Center for Judicial Studies. Reprinted with permission. © 2018 Duke University School of Law. All rights reserved. JUDICIALSTUDIES.DUKE.EDU/JUDICATURE

DATA VALIDATION

A crucial step toward controlling and understanding your data

BY GEORGE SOCHA AND SAAYA SHAH

WE ALL KNOW THE VOLUME OF DATA IN LITIGATION – PARTICULARLY EMAIL DATA – CONTINUES TO GROW RAPIDLY, WITH NO SIGN OF ABATING. That growth is forcing litigants to come up with ever better ways of quickly identifying and setting aside nonrelevant messages and finding messages of significance as early as possible.

These days litigants regularly use techniques such as deduplication and de-NISTing to reduce the amount of data that needs to be reviewed. Deduplication is the process of identifying identical copies of files; if two or more copies of a file have been identified, one copy is kept in the discovery set and the other copies are set aside. Often this process is conducted using what's called a cryptographic hash function — a mathematical algorithm that assigns a hexadecimal number (or a hash value) to each file. Two files with the same hash value almost always are identical files.

De-NISTing is a specialized form of deduplication. The National Institute of Standards and Technology (NIST) maintains the National Software Reference Library (NSRL). The NSRL contains a list of hash values (unique numeric identifiers for data) for known software files, such as operating system and application files, referred to as the NIST List. NIST List files rarely are of any interest in law suits or investigations. Because of this, e-discovery providers, law firms, and others can run the NIST List against discovery documents to cull files with hash values matching those in the NIST List to reduce the total volume of data.

Well-executed deduplication efforts like these can significantly reduce the volume of data subject to further analysis or review — though by how much depends on specific circumstances. In a 2012 Rand report, "Where the Money Goes: Understanding Litigant Expenditures for Producing Electronic Discovery," vendors estimated the percentages of near-duplicate documents in review as anywhere from 20-30 percent, 25-50 percent, 30-50 percent, and 30-60 percent. Actual results differ greatly depending on the characteristics of the document population as well as the methodologies used to identify duplicates and near-duplicates.[1]

Here, we'll focus on three additional techniques litigants can use to achieve the two goals of reducing data volume and finding key information: file type analyses, email domain analyses, and email timeline analyses.

All three techniques are forms of culling. File type analyses are used to identify the types of files in a population (.docx or .exe, for example). A common use of a file type analysis is to cull out — actually, set aside — files whose types are not likely to be of interest in a particular lawsuit or investigation. Email domain analyses are used to determine the domains from which or to which email messages are sent (gmail.com, for example). Email domain analyses can be used to identify email domains for messages not likely to be of interest. Email timeline analyses, while more complicated to perform, can achieve the same basic objective: to identify files outside the scope of what is important for a particular matter.

Litigants may use all three techniques and more to carve off large chunks of data that no longer need to be assessed for responsiveness, privilege, and the like and to home in more quickly on the data that matters. These various tools can be used in whatever order makes the most sense for the project at hand.

▸

FILE TYPE ANALYSIS

File type analyses start with two key sets of information: the number of electronic files collected (if they are your client's files) or produced (if they come from another party), and the types of those files (email, word processing, etc.). Typically this information is obtained by using some form of e-discovery processing tool.

At this point a word (or two) of warning is appropriate: Because each tool processes data differently, there is a very real chance that two different tools starting with the same set of data will deliver two different sets of results. File counts could differ because of different approaches taken to define what is and what is not a document. Consider the example of a PowerPoint presentation containing an embedded Excel spreadsheet. Is that one file, or two? The number of files of any particular file type also could differ. This is because there are at least two ways of identifying file types. The types of files in a population could be determined by looking at the extensions shown for the files, such as ".docx" for "data-validation-article.docx." This approach seems easy but does not always produce reliable results. You could, for example, manually change the extension at the end of the file from ".docx" to ".xyz", so that the new name would be "data-validation-article.xyz." Then the file no longer would be counted as a Microsoft Word file. A better approach is to use specialized processing software that analyzes the contents of files and compares those contents with the extensions appearing in the file names; this process will identify discrepancies such as the one created by changing a file name extension.

Once file types are obtained, you can analyze your data in many different ways. A chart or graph provides an at-a-glance overview of the composition of the population. Figure 1 shows that email messages make up the majority of files collected in the sample (.msg files account for nearly 60 percent of the population). Chances are, you will want to make sure files of this type go into the "keep working with these" bucket rather than in the "set aside and probably do not to look at again" bucket.

This could be a case where audio matters. If so, you would want to look at the list of file types to the right of the chart. You might notice it includes five types of files that almost certainly contain sound (".mpe", ".wav", ".mpg", ".mp3", and ".mpeg") as well as other file types that may have sound (such as ".ppt" files).

FIGURE 1. DOCUMENT EXTENSIONS – CHART VIEW

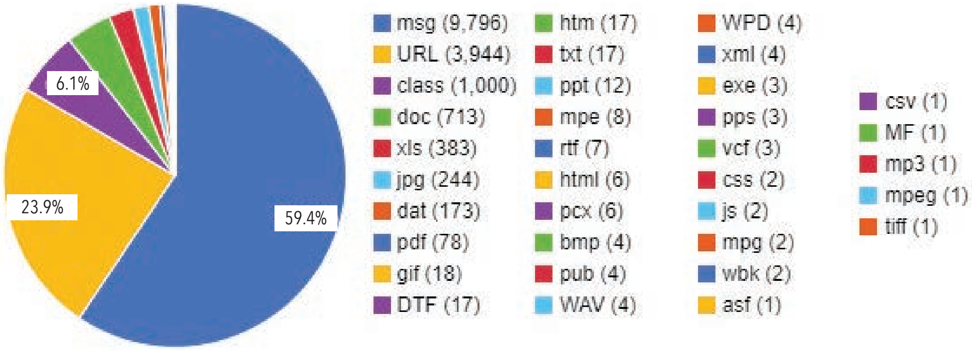
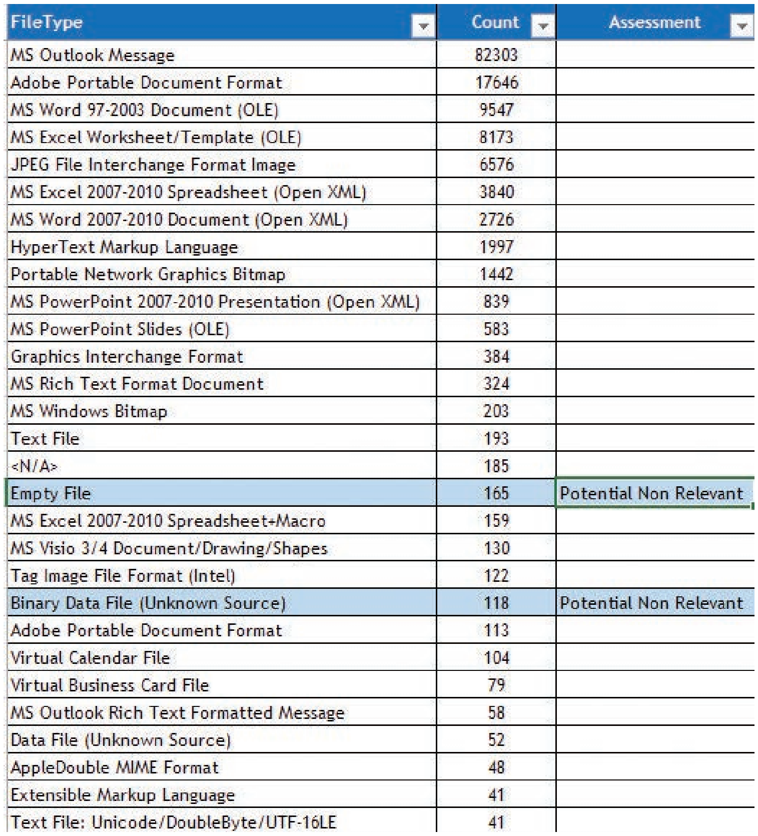


FIGURE 2. DOCUMENT EXTENSIONS SORTED – SPREADSHEET



FileType	Count	Assessment
MS Outlook Message	82303	
Adobe Portable Document Format	17646	
MS Word 97-2003 Document (OLE)	9547	
MS Excel Worksheet/Template (OLE)	8173	
JPEG File Interchange Format Image	6576	
MS Excel 2007-2010 Spreadsheet (Open XML)	3840	
MS Word 2007-2010 Document (Open XML)	2726	
HyperText Markup Language	1997	
Portable Network Graphics Bitmap	1442	
MS PowerPoint 2007-2010 Presentation (Open XML)	839	
MS PowerPoint Slides (OLE)	583	
Graphics Interchange Format	384	
MS Rich Text Format Document	324	
MS Windows Bitmap	203	
Text File	193	
<N/A>	185	
Empty File	165	Potential Non Relevant
MS Excel 2007-2010 Spreadsheet+Macro	159	
MS Visio 3/4 Document/Drawing/Shapes	130	
Tag Image File Format (Intel)	122	
Binary Data File (Unknown Source)	118	Potential Non Relevant
Adobe Portable Document Format	113	
Virtual Calendar File	104	
Virtual Business Card File	79	
MS Outlook Rich Text Formatted Message	58	
Data File (Unknown Source)	52	
AppleDouble MIME Format	48	
Extensible Markup Language	41	
Text File: Unicode/DoubleByte/UTF-16LE	41	

By contrast, more than 6 percent of the files in the chart are "class" files. A "class" file is something meant to be run by a computer rather than read by a human. For most lawsuits, "class" files do not contain content bearing on any disputed issues. Those files can be set aside, further reducing the volume of data needing review.

It also helps to put file type information into a tool where it can be sorted, filtered, and marked up. Figure 2 is an example from an Excel spreadsheet.

This example contains three columns: file type, which lists the types of files in the collection; count, which shows how many of each type of file were found; and assessment, where comments such as "Potential Non Relevant" have been added. The contents are sorted from largest count to smallest, so that at the top of this list is "MS Outlook Message" with 82,303 files and at the bottom (not shown) are "MS PowerPoint 2007-2010 Presentation+Ma" and "MS Word 2007-2010 Document-Macros (Open XML)" with 15 each. The contents could be sorted in other ways, as by FileType in ascending alphabetical order.

Because the file type information is in a spreadsheet, it can be filtered. You could set a filter to display only those rows where "Excel" appears in the "FileType" column. If you did that, you would see three rows containing information about Excel files:

FILETYPE	COUNT
MS Excel Worksheet/Template (OLE)	8,173
MS Excel 2007-2010 Spreadsheet (Open XML)	3,840
MS Excel 2007-2010 Spreadsheet+Macro	159

Filtering by file type allows you to restrict the data to be analyzed, processed, or otherwise handled to only those file types most likely to be of significance to your matter. You might want to focus first on Office files. If so, you could filter for file extensions such as ".doc" and variants such as ".docx" as well as ".ppt", ".xls", and their variants. Or you might decide that you had no interest in looking at graphics files, in which case you might filter to exclude such file types as ".bmp", ".jpg", and ".png".

File type filtering frequently is used to eliminate picture, video, and music files, which often are personal rather than business files. Eliminating these types of files can effectively reduce the overall volume of data for review, even when the number of eliminated files is small, because audio, video, and image files tend to be much larger than most other types of files.

Filtering can be done using more than one criterion. If you had the requisite information available, you could, for example, filter to see how many email messages were received by Custodian X between Jan. 1 and Feb. 15, 2016.

Finally, you could mark up the spreadsheet containing file type information. In the example in Figure 2, the comment "Potential Non Relevant" has been added in the "Assessment" column for file types that were deemed to be types not likely to contain any relevant information: "Empty File," "Binary Data File (Unknown Source)," "Thumbs Plus Database," and "Source Code (General)."

EMAIL DOMAIN ANALYSIS

Email domain analysis looks at the domains that email messages are sent from or to. Filtering by email domain, especially by the domain from which emails are sent, can be a quick and effective way of identifying large blocks of email messages that are likely irrelevant, potentially privileged, or presumptively responsive.

An email domain is the part of an email address that comes after the @ symbol. In the email address "updates@fantasyfootball.com," for example, the domain is "fantasyfootball.com." The domain is a part of the header information that accompanies every email message. As long as email messages handled in discovery are kept in native or near-native forms, such as ".msg" or ".eml", the email domains are readily available. If email messages are converted to image formats such as ".pdf" or ".tif" or printed to paper, the domains might become harder to access or might no longer be available.

Examining email domains can be a swift and convenient way to identify and set aside communications that are unlikely to have any meaningful connection with the dispute at hand. These include such things as emails from shopping, news, and sports websites; social media notifications; newsletters; digests and other mailing alerts; and spam and phishing messages.

Email domain analysis also can be a cost-effective way of tracking down communications such as email messages from other parties in a lawsuit, emails from other organizations of special interest, or messages sent by members of law firms where the messages might contain privileged communications.

Email domain information can be presented in charts, graphs, tables, and various other formats. A pie chart, using data from the EDRM Enron data set, shows the top ten email domains from which a set of email messages were sent (see Figure 3, next page). Almost 90 percent of the email messages came from the "enron.com" domain, which is no surprise. Also in the top ten, however, are messages from a newspaper (nytimes.com); a college (williams.edu); and a mass e-mailing system (mailblaster.clearstation.com).

EMAIL TIMELINE ANALYSIS

Email timeline analysis focuses on the dates and times of email messages and similar materials such as calendar entries. Five fields of metadata are especially useful for this type of analysis:

- Date and time email was sent;
- Date and time email was received;
- Date and time email was created;
- Date and time email was last modified or was deleted; and
- Appointment start and end dates and times.

That information can be combined with other content available from or about the messages, such as individual senders or recipients, types of files attached to messages, and email domains, to create a variety of timelines that can help you identify gaps, times of heightened activities, and any number of trends.

Figure 4 is an example of a visual timeline that shows the total numbers of email messages sent by several custodians in seven years (1979, 1980, 1998, 1999, 2000, 2001, and 2002).

FIGURE 3. SAMPLE EMAIL DOMAIN ANALYSIS – ENRON DATASET

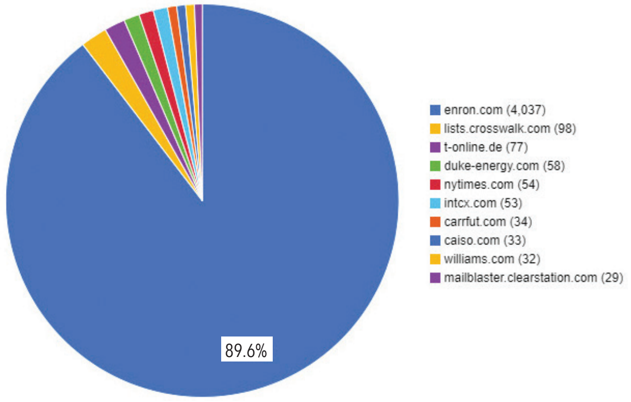
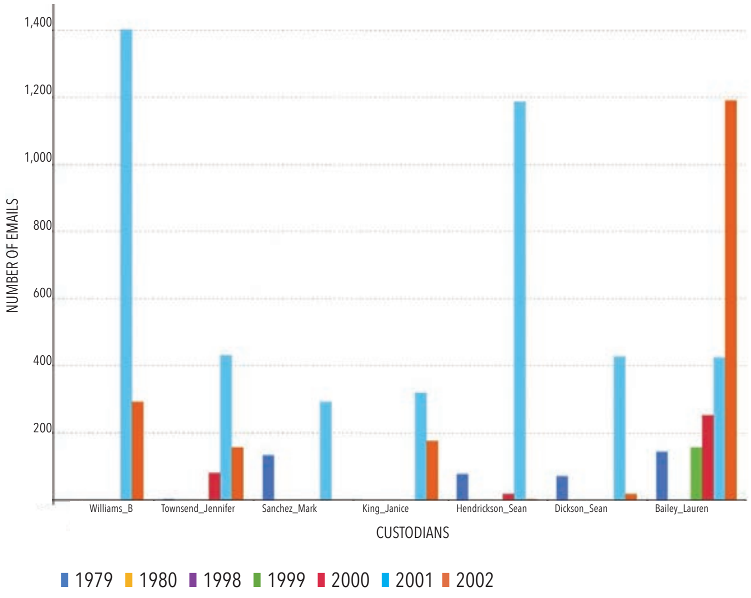


FIGURE 4. SAMPLE EMAIL TIMELINE ANALYSIS



Looking at the third custodian, we see he sent messages in 1979 and in 2001, but did not send any in the intervening years. This gap — the missing years — might be something we want to examine more closely.

Often these types of visual timelines are dynamic. You might be able to focus in on a narrower timeframe, with the example zooming in on just one year and then displaying counts by month rather than by year. You might be able to incorporate other pieces of information, such as domain names. You might reorganize how the data is displayed, so that the bottom axis is organized by years instead of by custodians.

DATA VALIDATION CAN BE a powerful tool both for carving off large chunks of data that no longer need to be assessed and for homing in quickly on data that matters greatly for issues at hand. Deployed effectively, these three forms of data validation — file type analysis, email domain analysis, and email timeline analysis — and similar tools can help litigants contain the costs of e-discovery and sharpen their focus on information that can help bring matters to resolution more quickly.

– **GEORGE SOCHA** *is managing director at BDO and co-founder of EDRM, an organization of e-discovery professionals that is now part of the Bolch Judicial Institute of Duke Law School.* **SAAYA SHAH** *is a senior manager at BDO.*

[1] Nicholas M. Pace, Laura Zakaras, *"Where the Money Goes: Understanding Litigant Expenditures for Producing Electronic Discovery*," Rand Corp. (2012), https://www.rand.org/content/dam/rand/pubs/monographs/2012/RAND_MG1208.pdf., at 52–53.