The International Journal of Digital Curation
Issue 1, Volume 6 | 2011

Assessing the Preservation Condition of Large and Heterogeneous Electronic Records Collections with Visualization

Maria Esteva, Weijia Xu, Suyog Dutt Jain, Texas Advanced Computing Center

Jennifer L. Lee, Wendy K. Martin,

University of Texas Libraries,

University of Texas, Austin

Abstract

As collections become larger in size, more complex in structure and increasingly diverse in composition, new approaches are needed to help curators assess digital files and make decisions about their long-term preservation. We present research on the use of interactive visualization to analyze file characterization information for the purpose of assessing the preservation condition of a vast collection of complex electronic records. The case study collection contains over 1,000,000 files of diverse formats arranged in varied record structures and record groups. The visualization application uses tree maps and a relational database management system (RDBMS) to represent the collection's arrangement and to show available characterization information at different levels of aggregation, classification and abstraction. Through this visualization interface curators can interact dynamically with the collections' characterization information to discover trends, as well as compare and contrast various file characteristics across the collection. Curators may select and weight the variables that they want to analyze. They can pursue analysis workflows that go from a high-level overview of the collection's preservation condition based on file format risks, to obtaining more detailed results about the condition of record groups and individual records. While there are various digital preservation planning tools available, to our knowledge none have been designed specifically to visually present assessment information across vast and complex collections. We present research to address the need for such a tool.[1]

[1] This paper is based on the paper given by the authors at the 6th International Digital Curation Conference, December 2010; received December 2010, published March 2011.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. ISSN: 1746-8256 The IJDC is published by UKOLN at the University of Bath and is a publication of the Digital Curation Centre.

Introduction

In this research, we investigate the use of visualization to aid and enhance the preservation assessment of very large and complex electronic records collections.

Preservation assessment of electronic records collections is a multi-layered process, the analysis of which is unique to each collection. Multiple variables must be considered in order to arrive at decisions about how to maintain collection accessibility over time (Boyle & Humphreys, 2008), and digital curators are often confronted with challenges and unknowns throughout the evaluation process.

A fundamental piece of preservation assessment is file format characterization. Characterization includes identifying file formats and ascertaining the preservation risk factor associated with those files based on internal institutional policies and/or established sustainability criteria (JHOVE2, 2010; PLANETS, 2010). However, the characterization step may not offer a complete picture of a collection's preservation condition. For example, a collection may contain digital objects with file formats that are not identifiable, files that have not been evaluated in terms of sustainability or files for which there is no further information beyond basic format identification. Learning what is not known about a collection is an important part of its assessment.

Equally important is consideration of the arrangement of files within a collection. An electronic record may be formed by more than one file and therefore may contain various file formats bearing different preservation conditions.[2] In turn, records may be arranged in directory structures that are key to their functionality and understanding, and a collection may have multiple types of records and groups of records. Ideally, the preservation assessment should be based on records, but in many cases identifying what constitutes a record can be challenging, as it may depend upon institutional policy, the functional capabilities of a particular set of files, or it may be derived by the user (Duranti & Thibodeau, 2006).

The criteria on which different institutions base their assessment and their preservation decisions are factored into the analysis. Institutions establish priorities considering variables such as the size of the collection, the presence of certain classes of data (Cornell, 2004), and the known risk level of the collection (Pardo et al., 2006). Comparing and contrasting the condition of different groups of records or of different collections is also a common way of prioritizing preservation decisions.

As collections become larger, the challenges of analyzing characterization data are compounded (Frost, 2009). To address the issues present in large electronic records collections, we developed a proof of concept visualization to analyze characterization information of a collection of 1,031,118 files. Using the visualization, curators can explore characterization data interactively and in context with the collection's structure to identify groups of records. Curators can go from overviews to detailed examinations, combine variables of interest, identify risk priorities, detect patterns and make preservation recommendations based on the assessment.

[2] In this context, a record can be considered a complex data object.

Related Work and Discussion

Visualizations allow users to interact with, infer knowledge from, and make decisions about large and complex data sets (Thomas & Cook, 2005). Work in the areas of business, health care and bioinformatics offer examples of the use of visualizations to tackle domain-specific problems by integrating principles and research methods used in the particular field into a visualization framework (Weber et al., 2009; Brownstein et al., 2010; Rudolph et al., 2009). Specifically in the area of archives, visualizations have been developed to explore finding aids (Kramer-Smyth, 2009; Whitelaw, 2009) and to observe personal electronic record-keeping practices (Xu et al., 2010). To the extent of our knowledge, this is the first visualization to aid in interactive assessment of the preservation condition of digital collections.

It is appropriate to understand this research vis à vis recent developments in digital preservation planning and management tools. Project PLANETS is a planning framework that integrates characterization services with preservation recommendations and preservation actions (PLANETS, 2010). In turn, JHOVE2, which has not been released at the time of this paper, will characterize simple and complex digital objects of a number of supported file types and enable their assessment against policy-based rules (JHOVE2, 2010). Our project differs from and can potentially complement both. It is a tool to explore large sets of characterization information, including such data as generated by the aforementioned projects. Its main features are interactivity and data integration. As opposed to evaluating discrete characterization variables, a curator can combine variables dynamically. Results are presented visually to obtain unified views that facilitate comprehension.

Visualization Design and Implementation

Requirements

We considered the following requirements in the visualization design:

1. Characterization data should be presented to show what is known and what is not known about the collection;
2. The collection's structure should be represented to observe relationships between data objects;
3. It should allow comparing and contrasting collections or groups of records for purposes of identifying patterns and establishing priorities;
4. Curators should be able to obtain high level and detailed preservation condition view on the fly;
5. Curators should be able to filter characterization variables and change their weight to reflect their criteria in the analysis results;
6. Curators may promptly identify problem areas in the collection as well as areas in good preservation condition.

Design Process

Our design team includes data analysis and visualization experts, archivists, and preservation librarians. After discussing the visualization requirements we followed an iterative design process in which each of the application's versions was evaluated by the team. In turn, the feedback was incorporated into the design and the modifications were discussed further. Design discussions focused on the function of the different preservation views and statistics gathering, visual representations through color, shapes and layout, and evaluation of the consistency between views of preservation condition at the collection level and more detailed ones at the record group level.

Case Study

As a case study, we used a portion of the collection in CI-BER (Cyberinfrastructure for a Billion Electronic Records), a research testbed developed by the National Archives Center for Advanced Systems and Technologies (NCAST). The electronic records are provided by federal agencies or harvested from their web sites. The collection does not have a finding aid and is organized in 125 records groups, each belonging to a different federal agency. In turn, each record group may have more than one subgroup of records.

Treemaps

Treemaps are a space efficient method to visualize large amount of hierarchical data simultaneously, where each directory is represented by a rectangle and all its subdirectories in the form of nested rectangles within it (Bederson et al., 2002). For each directory, the characterization information is rendered on its corresponding rectangle. This provides the user with a single combined view of the characterization information across the collection and allows the curator to compare and to identify patterns.

Metadata: Classification and Abstraction

To represent how files are logically arranged in groups, we use the collection's structural metadata, which includes the files' paths and the files' sizes. To gather this metadata from the collection's storage system, we traverse through the hierarchically arranged data and store the metadata in the relational database management system (RDBMS).

We derive characterization metadata from: a) file format identification information extracted with DROID, (DROID, 2006) and b) sustainability scores from the Stanford Digital Repository Format Scoring Matrix (Anderson et al., 2005).[3] In the Stanford criteria, sustainability scores indicate the preservation risk of a file format and their values range from zero to five. Since DROID does not recognize every file format present in the collection, and only a subset of those identified has an assigned sustainability score, one of our goals was to show how such missing information influences preservation assessment.[4] Currently, we have 1,031,118 files in our database, 90% of which were identified by DROID, with 200 different file formats. To summarize the large amount of file format information generated for this collection,

[3] While for purposes of this proof of concept we use an existing risk scoring criteria, the visualization may be adapted to assess characterization information derived from other methods such as JHOVE or PLANETS.

[4] The file formats with assigned scores include: plain and marked up text, tiff, jpg, png, bmp, photoshop, pdf, office docs, flash, wave, aiff, mp3, mpg, real, qt and windows media.

we classified file formats into 22 classes according to their functions.[5] In the RDBMS, metadata are aggregated at the directory levels and different statistics can be generated and rendered visually on demand as the user interacts with the visualization.

Visualization for Preservation Assessment

Because it can be difficult to comprehend large amounts of data all at once, well-designed visualizations should provide analysis paths that lead to a clear understanding of massive amounts of data in a manner that goes from an overview to a detailed view (Stuart et al., 1999). As with any software, using the visualization involves a period of adjustment to the application, in which users learn the visual metaphors and the kinds of analysis enabled by the application. Assessment is conducted at the users' pace, allowing them to integrate their experience to the analysis.

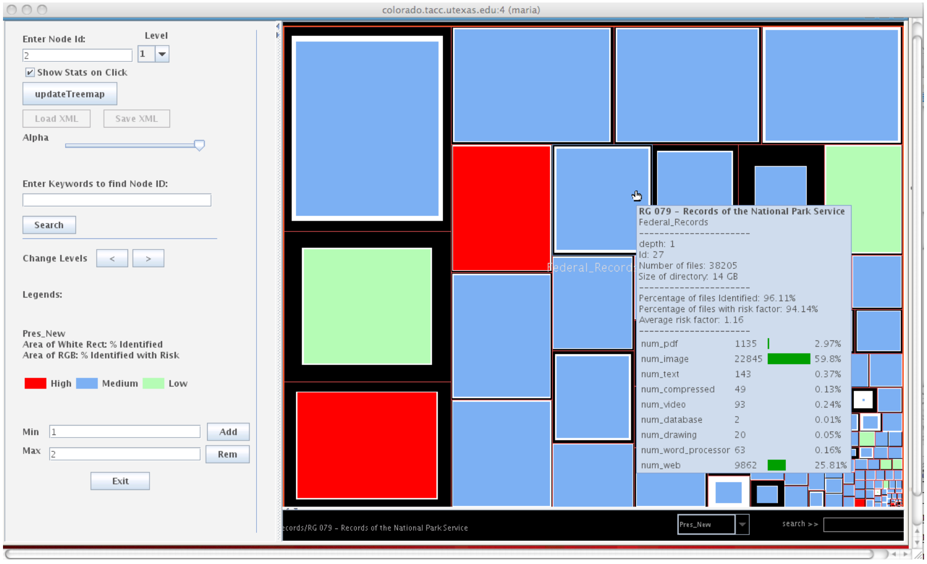
High Level Collection Preservation View

Figure 1 is a view of the entire collection showing high-level preservation information. Each square (surrounded by red boundary lines) represents the top-level directory of a record group, and the number of files within determines the size of the square. Therefore, the larger the square, the more files present within that record group. For each directory, different sizes, shapes and colors are used to represent preservation statistics. The outer black area represents the percentage of files whose format is unknown, whilst the white area represents the percentage of files that have been identified but do not have sustainability scores.[6] The size of the inner most-square is proportional to the percentage of files that have been identified and have a sustainability score. Its colors, whose corresponding values are shown in the control interface at the left of the screen, correspond to a coarse assessment based on average sustainability scores for files in the directory. Sustainability scores are further summarized into three risk levels according to Stanford's criteria in which green represents low risk (a sustainability score of 0-1), blue represents medium risk (2-3), and red represents high risk (4-5). Upon pointing with the mouse to any directory of interest, a tool tip shows the record group or directory name, its position in the hierarchical structure and general statistics.[7] By highlighting the general condition of all the record groups, the collection preservation view allows the comparison of conditions between record groups, identification of those at higher risk, determines what is not known about each and guides the curator into the next steps of the assessment.

[5] Classes are: images, audio, GIS, database, web, word-processor, spreadsheet, PDF, drawing, text, video, XML, compressed, flash, publishing, graphics, email, print, development, OCR, calendar, and schema. Some classes contain only one file format, including all the versions currently identified by DROID.

[6] The use of areas within the squares to represent different numeric values was drawn from Mitchell Whitelaw's Series Browser (2009)

[7] Statistics are: total number and total size of files, percentage of files identified, percentage of files with a sustainability score, and the average sustainability score for the directory.



RG 079 – Records of the National Park Service
Federal_Records
----
depth: 1
Id: 27
Number of files: 38205
Size of directory: 14 GB

Percentage of files Identified: 96.11%
Percentage of files with risk factor: 94.14%
Average risk factor: 1.16
----

num_pdf	1135	2.97%
num_image	22845	59.8%
num_text	143	0.37%
num_compressed	49	0.13%
num_video	93	0.24%
num_database	2	0.01%
num_drawing	20	0.05%
num_word_processor	63	0.16%
num_web	9862	25.81%

Figure 1. High Level Collection Preservation View.

While the averages and the correspondent low, medium and high-risk color representation provide a summary of the condition of a directory, distributions generated on the fly provide details of the summarized condition. As shown in Figure 2 below, by clicking with the mouse on any directory, four types of pie charts are shown: a) the distribution of classes of files, b) the total distribution of risk levels, c) the distributions of classes of files per risk level, and d) the distribution of classes of files with no sustainability score. In addition, the curator may be interested in learning what file types correspond to which classes in the observed directory. This information can be retrieved from the database, if needed.

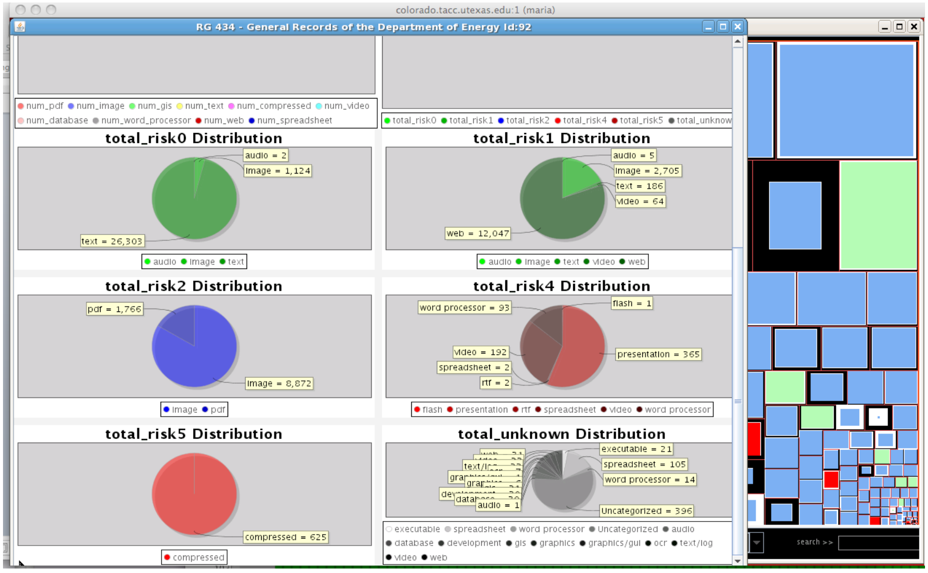

colorado.tacc.utexas.edu:1 (maria)
RG 434 - General Records of the Department of Energy Id:92
● num_pdf ● num_image ● num_gis ● num_text ● num_compressed ● num_video ● num_database ● num_word_processor ● num_web ● num_spreadsheet
● total_risk0 ● total_risk1 ● total_risk2 ● total_risk4 ● total_risk5 ● total_unknow
total_risk0 Distribution
audio = 2
image = 1,124
text = 26,303
● audio ● image ● text
total_risk1 Distribution
audio = 5
image = 2,705
text = 186
video = 64
web = 12,047
● audio ● image ● text ● video ● web
total_risk2 Distribution
pdf = 1,766
image = 8,872
● image ● pdf
total_risk4 Distribution
word processor = 93
flash = 1
video = 192
spreadsheet = 2
rtf = 2
presentation = 365
● flash ● presentation ● rtf ● spreadsheet ● video ● word processor
total_risk5 Distribution
compressed = 625
● compressed
total_unknown Distribution
executable = 21
spreadsheet = 105
word processor = 14
Uncategorized = 396
audio = 1
○ executable ● spreadsheet ● word processor ● Uncategorized ● audio ● database ● development ● gis ● graphics ● graphics/gui ● ocr ● text/log ● video ● web
search >>

Figure 2. Distributions Per Record Group in the Collection Preservation View.

Interacting with Characterization Variables

We developed two interactive views—the aggregator and the selector—that allow curators to select, combine, and rank or filter characterization variables to explore preservation scenarios and aid in establishing priorities. The variables are the statistics generated from the characterization data (size and number of files, file classes and levels of risk, including unknown risk), which are aggregated in the RDBMS at the record group level. Variables can be analyzed individually or they can be combined to learn which ones are present across a given record group. Figures 3 and 4 show the aggregator and the selector views respectively.

The aggregator view is useful in discovering conditions within a collection. Using the control tool, curators can assign different weights to each of the variables that are relevant to their preservation criteria. The system computes the weighted average of the selected characterization variable for each record group and renders ranked results divided into five percentiles, which are represented with colors. The highest ranking record groups are highlighted in pink, and the rest in shades of yellow, from bright (high ranking) to pale (low ranking). If none of the selected variables are present in a record group, the directory remains black.

The selector view allows curators to select one or more variables and apply constraints on their values. The system then highlights in white only those record groups that satisfy all the constraints applied in a specified order, showing in black those that do not. Both views allow the comparing and contrasting of conditions between record groups. For purposes of illustration, we chose to combine some of the most problematic variables and assigned the same maximum weight to all in both the aggregator and selector views. Those were a) risk level 4, b) risk level 5, and c) files with unknown scores. In the aggregator view we can observe that, with the exception of four small record groups shown in black, the record groups aggregate at one or more of the variables in different degrees. The selector view presents only the record groups that combine all three variables. We concluded that record groups highlighted in pink in the aggregator view and in white in the selector view are the most problematic according to the criteria established in this example.

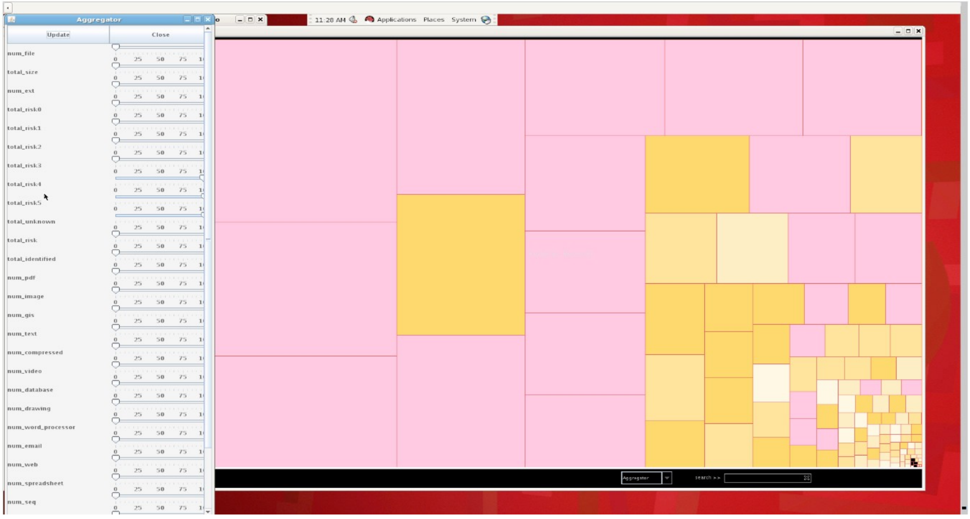
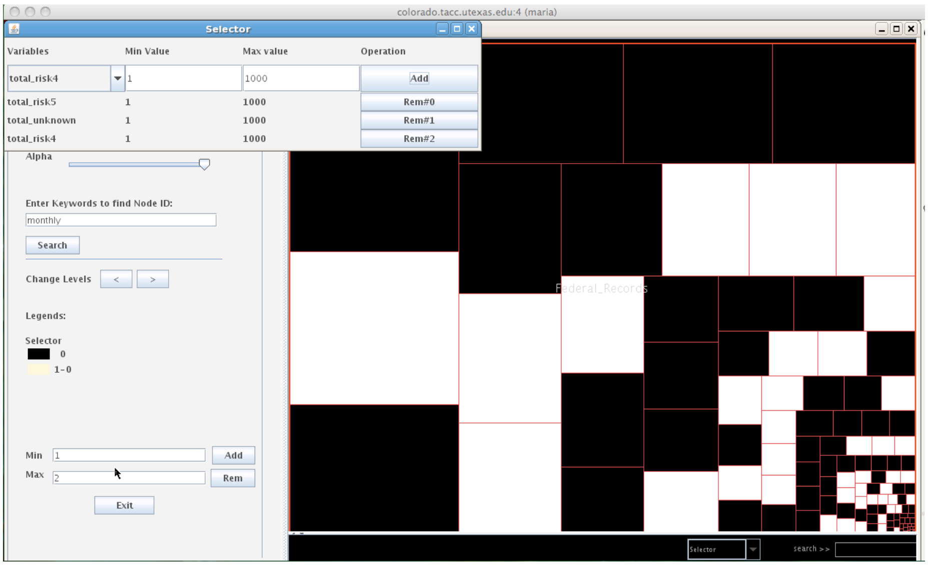


Figure 3. Aggregator view with variable and weight selector to the left.



Variables	Min Value	Max value	Operation
total_risk4	1	1000	Add
total_risk5	1	1000	Rem#0
total_unknown	1	1000	Rem#1
total_risk4	1	1000	Rem#2

Enter Keywords to find Node ID: monthly

Search

Min: 1

Max: 2

Selector

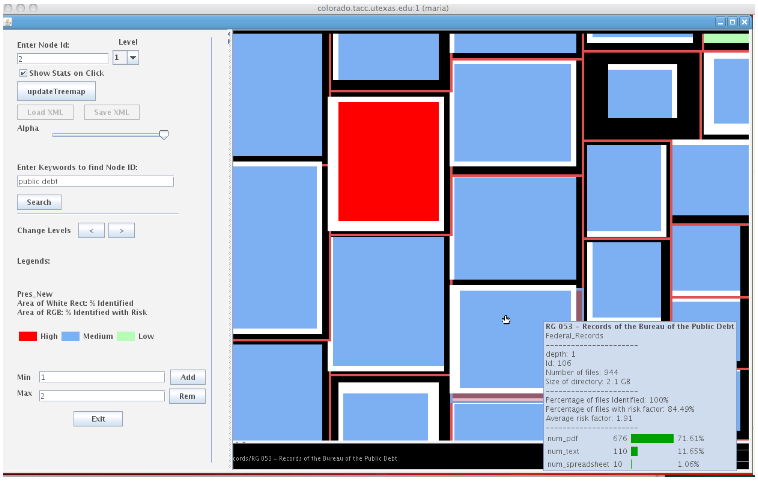
Figure 4. Selector view with variable filter in the top left side.

So far, we have shown preservation views based on individual file characterization information. In the following section we show the possibility for the assessment of preservation condition at the record group and/or record level.

An Analysis Workflow: From Overview to Details

Record Group Preservation Assessment

We use the case of the record group “Records of the Bureau of Public Debt” to illustrate an analysis workflow that goes from a collection overview to a detailed examination of a record group and records. Figure 5 shows a zoomed view of a section of the entire collection in which the case study record group is shown, with its tool tip, in the collection context. From the tool tip we learn that this record group has a total of 944 files, all of which have identified file formats. The average risk level for files with a sustainability score is medium (2-3). However, the white rim in the border indicates that some of the files do not have a sustainability score.



Enter Node Id: 2

Level: 1

☑ Show Stats on Click

Enter Keywords to find Node ID: public debt

Pres_New
Area of White Rect: % Identified
Area of RGB: % Identified with Risk

Min: 1

Max: 2

```
RG 053 – Records of the Bureau of the Public Debt
Federal_Records
------------------------
depth: 1
Id: 106
Number of files: 944
Size of directory: 2.1 GB
------------------------
Percentage of files Identified: 100%
Percentage of files with risk factor: 84.49%
Average risk factor: 1.91
------------------------
num_pdf          676   71.61%
num_text         110   11.65%
num_spreadsheet  10    1.06%
```

cords/RG 053 – Records of the Bureau of the Public Debt

Figure 5. High-Level Preservation View of the Record Group in the Collection Context.

By clicking on the directory, distributions are generated on the fly. The pie charts presented in Figure 6 below show that this record group has four classes of file formats (spreadsheets, text, OCR and PDF), that the majority fall into the medium risk level (2-3), and that there are files without sustainability scores. In the total risk distribution pie charts it can be observed that the relationship between classes of files and risk level are one to one, and that the spreadsheet class has the highest risk score of 4. Since the OCR class shows an unknown risk score, we queried the database and found that they are OmniPage optical character recognition files. Though we did not assign a score to this file format, curators conducting analysis could do so to complete their assessment.

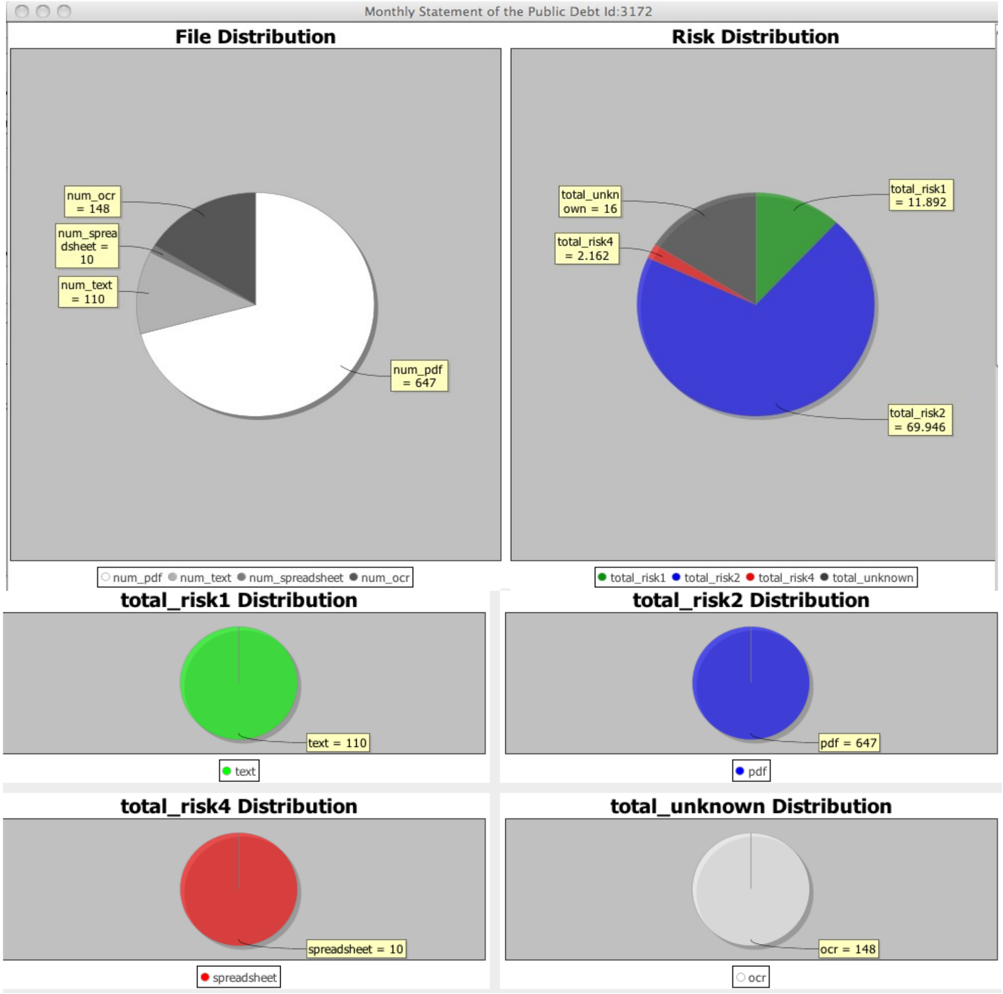
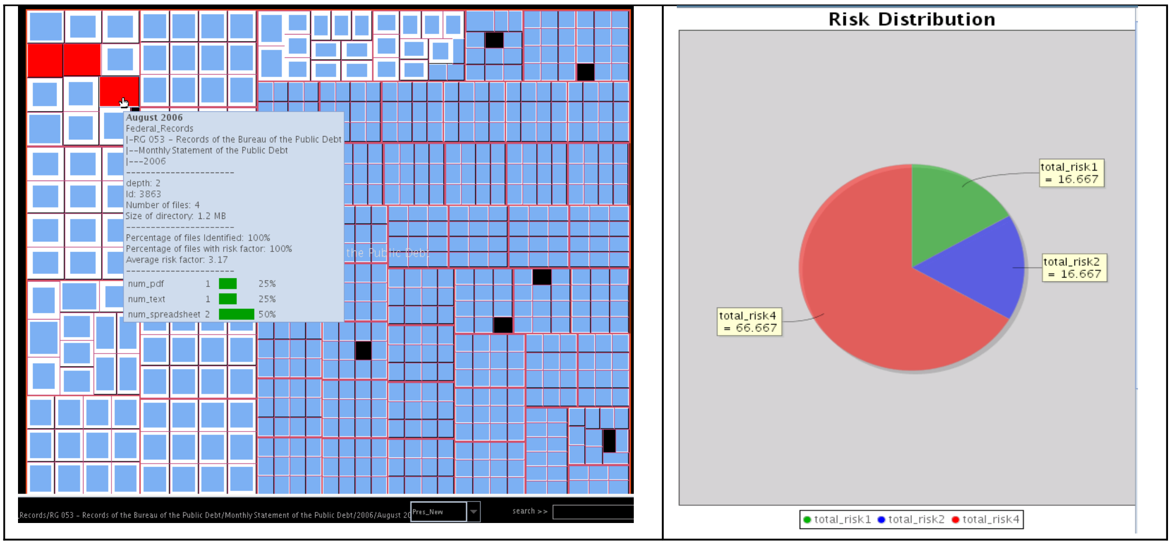


Figure 6. File Classes and Risk Distributions for the Case Study Record Group.

Using the control tool along the left side of the interface (shown in Figure 5), a curator can select the directory to explore and a view of that directory is generated. This detailed view allows the user to observe characterization information within the structure of the selected directory. The screenshot to the left in Figure 7 shows such a view for the case study record group, including all its subdirectories. In this process, characterization metadata is recalculated for each subdirectory showing: empty directories (black), directories with files at a medium risk level (solid blue), directories that include both medium risk files and files without sustainability scores (white outer areas and blue centers), and three high risk directories (red). In this view it is possible to identify the location of the directories at higher risk and their composition, as well as the directories for which some of the risk scores are unknown.[8] In turn, the screenshot to the right presenting the distribution of risk for one of the high-risk directories shows that the majority of the files in that directory correspond to high-risk files (in this case, spreadsheet files). Through the identification of patterns we can determine the similarities between the four types of subdirectories present in this record group.



```
August 2006
Federal_Records
|-RG 053 - Records of the Bureau of the Public Debt
|--Monthly Statement of the Public Debt
|---2006
----------------------
depth: 2
Id: 3863
Number of files: 4
Size of directory: 1.2 MB
----------------------
Percentage of files Identified: 100%
Percentage of files with risk factor: 100%
Average risk factor: 3.17
----------------------
 num_pdf          1   25%
 num_text         1   25%
 num_spreadsheet  2   50%
```

```
_Records/RG 053 - Records of the Bureau of the Public Debt/Monthly Statement of the Public Debt/2006/August 20
```

Figure 7. Detailed View of Risk Per Subdirectory in the Case Study Record Group.

Preservation Assessment Based on Records

Ideally, the high-level collection preservation view should be based on the condition of the records. However, given that determining records from files in the record group was not done a priori, we use the visualization of the records group structure to aid in identifying records and then recalculate their preservation condition. By examining the visualization and the content of a sample of files in storage, we determined that each subdirectory contains from one to four files, each one a version of a monthly statement in different file formats.[9] In the database each directory has an ID, so once the directories containing individual records are identified, those IDs can be retrieved from the database and entered back into a records table in the RDBMS, from which statistics and graphs are generated to reflect the preservation condition of the records.[10]

[8] As we mention, in this case the curator can assign a risk score to the OCR files and complete the assessment.

[9] We did not make a decision as to which version constituted the official record and, for testing purposes, considered the different files as components of one record.

[10] In cases of records formed by files included in various subdirectories, we retrieve the ID of the parent directory and obtain information about the record.

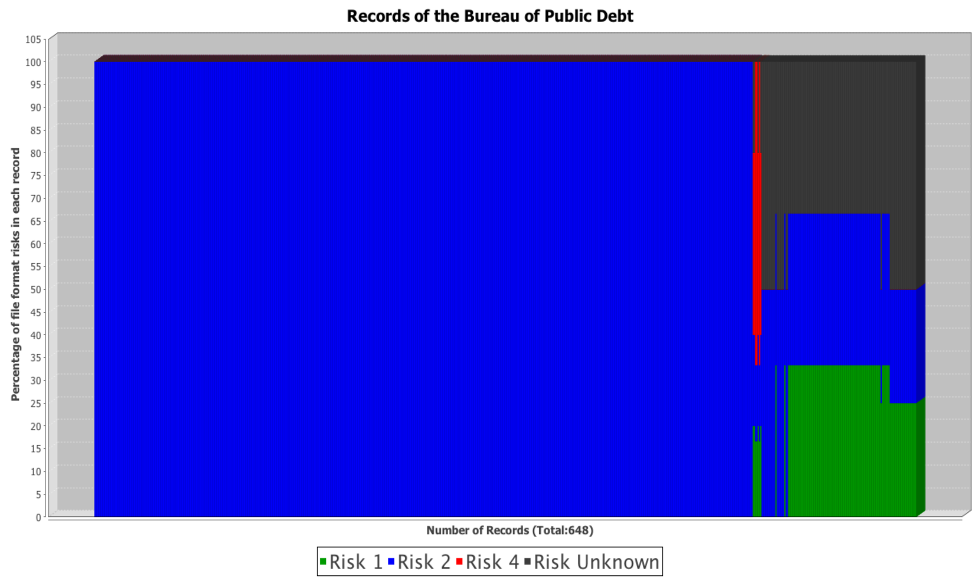


Figure 8. Stacked Bar Chart Showing the Preservation Condition of 648 Records.

In Figure 8, the stacked bar chart shows the percentages of risk levels in the 648 records in this record group, with each vertical bar representing one record. The bar's colors are based on the percentage of files at risk levels 1, 2, 4, and of unknown risk, which is represented in grey. The team evaluated the different views generated during this assessment and concluded that the transition between collection and records views are clear, and that the ability to identify file formats without sustainability scores facilitates focusing on them to seek further information. At the records level, ascertaining the risk of the individual files highlights problems in context and shows that there is enough information to make preservation recommendations. The team continues working to streamline the possibility of visually identifying records from files and to generate high-level preservation views based on the records condition.[11]

Conclusions

Assessment of preservation condition for large records collections is conducted in a case-by-case fashion, requiring significant amounts of time and expertise. It is a dynamic process in which different criteria may apply to different parts of a collection. In turn, due to its heuristic nature, the process is hard to model within a unified computational solution. This interactive visualization tool provides a flexible interface for curators to study characterization information derived from large and complex digital collections in an orderly way, with the possibility to combine multiple variables, and to apply their experience and institutional criteria to the assessment.

Acknowledgements

This work was supported through a National Archives and Records Administration (NARA) supplement to the National Science Foundation Cooperative Agreement (NSF) TERAGRID: Resource Partners, OCI-0504077.

[11] The team evaluated other cases of complex records. Identifying records from files and assessing their preservation condition is more straight forward for highly structured records/data objects.

References

Anderson, R. et al. (2005). The AIHT at Stanford University: Automated preservation assessment of heterogeneous digital collections. *D-Lib Magazine, 11, (12)*. Retrieved April 18, 2010, from http://www.dlib.org/dlib/december05/johnson/12johnson.html.

Bederson, B.B., Shneiderman, B., & Wattenberg, M. (2002). Ordered and quantum treemaps: Making effective use of 2D space to display hierarchies. *ACM Transactions on Graphics, 21, (4).* New York, NY: ACM.

Boyle, F., & Humphreys, J. (2008). Digital preservation planning: Principles, examples and the future with PLANETS. *Ariadne, Issue 57, October 2008*. Retrieved July 18, 2010, from http://www.ariadne.ac.uk/issue57/dpc-planets-rpt/.

Brownstein, J.S., Freifeld, C.C., Chan, E.H., Keller, M., Sonricker, A.L., Mekaru, S.R., & Buckeridge, D.L. (2010) Information technology and global surveillance of cases of 2009 H1N1 Influenza. *New England Journal of Medicine, 362.* Waltham, Massachusetts: Massachusetts Medical Society.

Cornell University Library. (2004). *Cornell University library digital preservation policy framework.* Retrieved August 5, 2010, from http://commondepository.library.cornell.edu/.

DROID, Version 4.0. (2006). The National Archives of the United Kingdom, PRONOM. Retrieved April 18, 2010, from http://droid.sourceforge.net/.

Duranti, L. & Thibodeau, K. (2006). The concept of record in interactive, experiential and dynamic environments: The view of InterPARES. *Archival Science 6 (1).* Netherlands: Springer.

Frost, H. & team (2009). Assessing digital objects with JHOVE2. *Presented at iPRES 2009.* San Francisco, CA. Retrieved July 20, 2010, from https://wiki.ucop.edu/display/JHOVE2Info/Project+Presentations.

JHOVE2 (2010). *The next generation architecture for format-aware characterization.* Retrieved January 4, 2011, from https://bitbucket.org/jhove2/main/wiki/Home.

Kramer-Smyth, J. (2009). ArchivesZ: Tackling the challenges of data aggregation. *Presented at the 2009 SAA Research Forum*. Retrieved April 18, 2010, from http://www2.archivists.org/proceedings/research-forum/2009/posters.Pardo, T.A, Burke, G.B. & Kwon, H. (2006). *Preserving state government digital information*. Center for Technology in Government. University of Albany, SUNY. Retrieved July 29, 2010, from www.digitalpreservation.gov/partners/pdf/ctg_dp_baseline2006.pdf.

PLANETS. *About PLANETS*. Retrieved April 18, 2010, from http://www.PLANETS-project.eu/about/.

Rudolph, S., Savikhin, A., & Ebert, D.S. (2009) "FinVis: Applied visual analytics for personal financial planning," Visual Analytics Science and Technology. *Presented at the IEEE Symposium on Visual Analytics Science and Technology (VAST).* Atlantic City, New Jersey.

Stuart, K., Jock K.C., Mackinlay. C.D., & Shneiderman, B. (1999). *Readings in information visualization: Using vision to think.* San Diego, CA: Academic Press.

Thomas, J., & Cook, K. (2005). *Illuminating the path: The R&D agenda for visual analytics.* National Visualization and Analytics Center. Retrieved April 18, 2010, from http://nvac.pnl.gov/agenda.stm.

Weber, G.H., Rübel, O., Huang, M. Y., DePace, A.H., Fowlkes, C.C.E., Keränen, Luengo Hendriks, C.L, Hagen, H., Knowles, D. W., Malik, J., Biggin. M.D., & Hamann, B. (2009). Visual exploration of three-dimensional gene expression using physical views and linked abstract views. *IEEE Transactions on Computational Biology and Bioinformatics, 2, (6).* Washington, DC: IEEE Computer Society.

Whitelaw, M. (2009). *The Visible Archive.* Retrieved April 18, 2010, from http://visiblearchive.blogspot.com/.

Xu, W., Esteva, M., & Jain Dott, S. (2010). Visualizing Personal Digital Collections. *Proceedings of the 10th annual joint Conference in Digital Libraries.* New York, NY: ACM.