Visualization for Archival Appraisal of Large Digital Collections

Weijia Xu, Maria Esteva, Suyog Jain Dott; Texas Advanced Computing Center, University of Texas at Austin; Austin, Texas, USA

Abstract

Our research examines data-driven visualization methods for archival purposes. Using data extracted from a large and heterogeneous digital collection, we created an information visualization that uses RDBMS and treemap to enable archival analysis. Different views present the collection's structure and properties at different levels of aggregation and abstraction, transforming 1,000,000 data points into information that enables observation and decision-making.

Introduction

Given the scale and diversity of digital collections, Archival Studies Professor Richard Cox states that there is a need to "look at a vast universe of documentation (and a universe that is expanding quite rapidly) and then to shrink it down in some strategic, planned, or rational fashion that allows archivists to administer it and researchers to access it" [1]. Digital archival appraisal is at the core of this problem, that is, analyzing collections to make decisions about their value and to determine their long-term retention needs. As digital collections grow in size and diversity, appraising them becomes more complex, and much more so if the collections do not have descriptive or technical metadata to begin with. While frameworks for conducting appraisal of digital collections exist [2], there are pressing needs for tools to facilitate the types of analysis and the decision-making processes that happen during appraisal.

Using relational database management system (RDBMS) and treemap visualization, we created an information visualization to enable archival analysis tasks in large digital collections. The visualization uses structural and technical metadata to represent large, heterogeneous, digital collections. It provides a series of functionalities based on data aggregations and categorizations to present the collection's properties interactively at different levels of abstraction.

Collection's Analysis[1]

Collections' analysis tasks for appraisal purposes include various forms of observation as well as decision making processes. Observations include determining the collection's scope and contents, the way in which it is organized, its completeness, technical characteristics, and its preservation condition. How digital objects are arranged, the file types and document types involved, and the relationship between them are factors that contribute to understanding the collection and its functions.

Traditionally, archivists evaluate collections by reading the records and or the labels that describe them; inspecting, counting, and browsing materials; and consulting indexes, inventories, and or catalogs when available. In most cases, the evaluation cannot happen at the object level and sampling methods are used to infer the total value and condition of the collection. Archivists also use secondary sources to investigate the collections' provenance and the reasons and methods through which the collection was formed. When there is not much information available about a given collection, archivists make informed inferences about its functions, uses, and the context in which it was created based on the observations and their professional experience. Between the evidence gathered during the analysis and what is inferred, archivists make decisions about the collection's value, its preservation condition, and its access and storage needs [3].

To address the scale and complexity of digital collections and to improve the precision and efficiency with which archivist conduct analysis, we propose using data driven methods. In this project, we use visualization to interpret the results of statistical summaries of large multivariate data. Our conceptual framework is based on investigative visual analysis, which focuses on presenting data at different levels of abstraction and enabling interactive exploration so users can derive insights from the data [4].

Case Study and Research Challenges

In this research we use the data collection in the Transcontinental Persistent Archives Prototype (TPAP), a research testbed developed by the National Archives Center for Advanced Systems and Technologies (NCAST) [5] to study the next generation of digital archival systems and services. TPAP uses iRODS as its grid data repository [6]. The records in the collection are publicly available digital records provided by Federal Agencies or harvested from their websites. The collection does not have finding aids and is organized by Record Groups (RG), each belonging to a different Federal Agency. In turn, each RG may have more than one sub-group bearing different arrangements and a variety of file formats and document types. The characteristics mentioned above, makes it an adequate testbed to investigate the following questions:

- What metadata can be extracted from large and multivariate digital collections?
- How can we visually represent this metadata in meaningful ways for archival appraisal purposes?
- What levels of abstraction allow making sense of a large and heterogeneous data collection without loosing perspective of the inherent characteristics of the individual sub-collections?
- What analysis methods are afforded by the availability of large amounts of data?

In this paper we show the visualization workflow, describe the metadata extraction processes and the database schema, present

[1] The discussion in this section is based on Jennifer Meehan's article "Making the Leap from Parts to Whole: Evidence and Inference in Archival Arrangement and Description," in which she discusses the way in which archivists analyze materials for arrangement and description purposes. The authors of this article considered that the analysis methods described also map with those conducted during appraisal.

the visualization's functionalities, and explain the way in which data is categorized, aggregated and presented to the user. We further discuss the types of analysis afforded by the visualization and whether they map appraisal tasks.

Workflow

Figure 1 below is a diagram of the visualization's workflow.

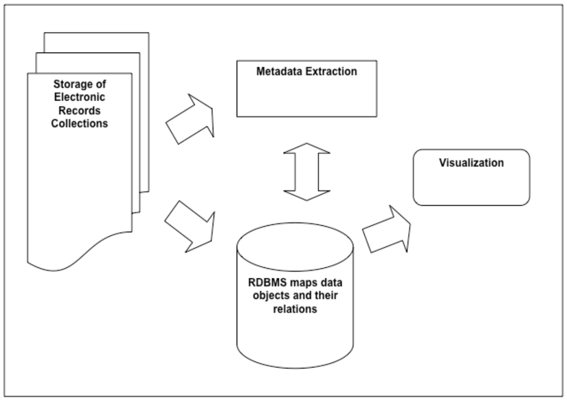

Storage of Electronic Records Collections
Metadata Extraction
Visualization
RDBMS maps data objects and their relations

Figure 1. *Visualization's workflow*

Metadata Extraction and Management

To visually represent the collection and its properties we use structural metadata extracted from the iRODS file system, and with file format identification metadata extracted from the files.

Structural metadata

We call structural metadata to the file system structure stored in iRODS, including name and paths of each directory and all the names and sizes of files within each directory. iRODS provides a Java API known as Jargon which contains methods that allow connecting to and querying the iRODS database through software applications. We wrote a Java client that establishes connection with the iRODS server. Once the connection is successfully established, we traverse through the hierarchical data in a Breadth First Search manner and aggregate metadata in a comma-delimited file. The resultant file contains the name, path, and size of the file in each row, which provides the basis for representing graphically the way in which files are grouped, and how groupings relate to each other in a collection.

File format identification metadata

In combination with the structural representation of the collection we use file format identification metadata to visualize the distribution of file formats in the different RGs. To extract file format identification information we use Digital Record Object Identification (DROID), a tool developed by the UK National Archives [7]. Currently we run DROID in a local copy of the testbed collection.

Data Categorization

DROID uses a signature file to match inspected files with their respective formats. We apply an XSLT to convert DROID's signature file data into an HTML table. This allows us to assign the file formats that DROID presently identifies to classes such as: images, audio, GIS, database, scripts, text, video, drawing, graphics, datasets, word processor, etc. This categorization allows abstracting large amounts of file format information to meaningful and manageable levels.

In addition, we assign the Stanford Digital Repository format score to each file format that has one [8]. The format score is based on the Sustainability Factors developed by the Library of Congress, against which file formats are assessed to determine their preservation feasibility [9]. Scores range from 0 (low quality value) to 5 (high quality value). From the score, we derive average preservation risk for a given directory that may include different file formats.

Table 1. *Technical metadata table*

Software	Version	PUID	Category	Score
Microsoft Word for Macintosh Document	6	x-fmt/2	word processor	4
Write for Windows Document	3.1	x-fmt/4	word processor	-
Works for Macintosh Document	4	x-fmt/5	word processor	-
FoxPro Database	2	x-fmt/6	database	-
FoxPro Database	2.5	x-fmt/7	database	-
AutoCAD Block Attribute Template		x-fmt/24	drawing	-
OS/2 Bitmap	1	x-fmt/25	graphics	0
JTIP (JPEG Tiled Image Pyramid)		fmt/149	image	1
JPEG-LS		fmt/150	image	1
JPX (JPEG 2000 Extended)		fmt/151	image	1
Waveform Audio (PCMWAVEFORMAT)		fmt/141	audio	0
Waveform Audio (WAVEFORMATEX)		fmt/142	audio	0
Waveform Audio (WAVEFORMATEXTENSIBLE)		fmt/143	audio	0

The Technical Metadata Table exemplifies how we assign the format categories and the format score to file formats identified by DROID. The limiting factor in this categorization is the incompleteness of the information. DROID does not recognize every file format present in the collection, and relatively few file formats have a format scoring assigned. In the preservation section we discuss how we consider this limitation and how the data is aggregated and presented to allow making about the collection's technical composition and sustainability. The output file generated by DROID in XML format is imported to the database.

Metadata Database

We store the technical and structural metadata using RDBMS, which serves as a centralized storage manager on disk and provides efficient data access that is scalable to the requirements of this project. The RDBMS has mature mechanisms to support concurrent access and provides low level data aggregation and transformation. We have developed code and procedures to integrate extracted metadata as well as to export it out from database in xml format for data sharing. The schema diagram of the relational database is shown in figure 2 below:

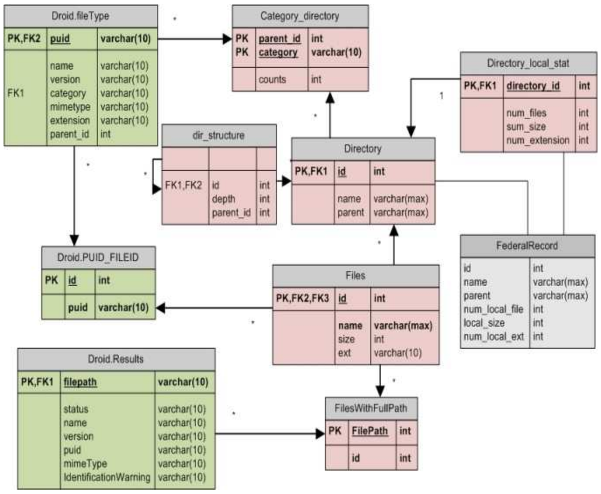

Droid.fileType
PK,FK2 puid varchar(10)
FK1 name varchar(10) version varchar(10) category varchar(10) mimetype varchar(10) extension varchar(10) parent_id int
Category_directory
PK parent_id int
PK category varchar(10)
counts int
Directory_local_stat
PK,FK1 directory_id int
num_files int sum_size int num_extension int
dir_structure
FK1,FK2 id int depth int parent_id int
Directory
PK,FK1 id int
name varchar(max) parent varchar(max)
FederalRecord
id int name varchar(max) parent varchar(max) num_local_file int local_size int num_local_ext int
Droid.PUID_FILEID
PK id int
puid varchar(10)
Files
PK,FK2,FK3 id int
name varchar(max) size int ext varchar(10)
Droid.Results
PK,FK1 filepath varchar(10)
status varchar(10) name varchar(10) version varchar(10) puid varchar(10) mimeType varchar(10) IdentificationWarning varchar(10)
FilesWithFullPath
PK FilePath int
id int

Figure 2. *Metadata database schema*

Currently in our database we have structural and file format information for 60,837 directories containing 1,031,118 files, 90% of which were positively identified and include 200 different file formats.

Visualization Features

As a form of representation we use treemap, a visualization developed by Ben Shneiderman in the late 1990's [10]. Treemaps are useful to present hierarchical structures as well as distributions of different types of data. We chose treemaps because we wanted to focus on structure as our primary way of representing the collection, and because it facilitates finding patterns and outliers. Due to the particularities of the archival tasks that that we wanted to support, we introduced various adaptations to the basic application. Our treemap required implementing database connectivity modules and usage of the Java Prefuse library [11]. It has basic functions that allow zooming in and out and repositioning the visualization. Upon pointing with the mouse to any directory, a tool-tip shows the directory's name, including the tree (hierarchical structure and level of nesting) information and statistics about the file categories present within.

Figure 3 below shows the collection represented as a treemap. Each square in the treemap represents one directory; the darker lines delimit the top-level directories corresponding to the 125 RG and the lighter ones the sub-directories within. The application employs various tiling algorithms and color schemes to achieve a visual representation of how the data is distributed in nested directories.[2] Using the control tool in the interface, users can specify the levels of nesting that they want to see and progressively observe the configuration of directories and sub-directories. The image below shows 4 levels of nesting from a total of 20. Notice how the different patters of arrangement emerge for each RG.

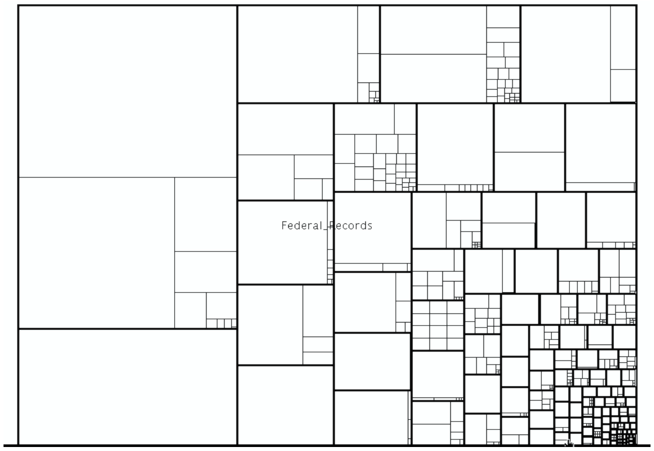


Figure 3. *Treemap visualization of the testbed collection*

In the collection's entire view, it is possible to distinguish the largest and the smallest RGs based on the number of directories within. The visualization also enables dynamic views of the total number and the total size of files within each directory at any given hierarchical level. Ranges of file numbers and sizes are represented by shades of color that go from light, (more file numbers or larger sizes) to dark, (less files or size density). These presentations facilitate making infrastructure, access, and preservation decisions.

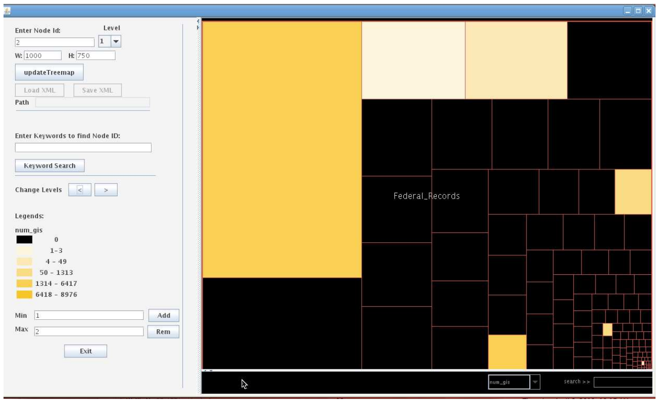
A main analysis method provided by the visualization is the possibility to learn by comparing and contrasting. In this way priorities can be established and decisions can be made based on what looks similar or different, what relates to what and what doesn't, and what is known and what is unknown.

Collection's Properties Views

Users can make selections in the visualization interface based on the stored metadata elements. These selections are aggregated at the directory level, mapped to different color values, and rendered on-demand. In Figure 4 below, which shows only the top-level directories (125), the shaded areas are directories containing GIS data, and the black areas do not contain any GIS data. In the control tool to the right, it is possible to see the correspondence between the different shades and the number of GIS files.

Views can be changed on the fly by selecting file type category options at any directory level. Using the entire view it is possible to identify all the RGs with a certain file format category. Beyond identifying the type of technologies included in the collection, these views can indicate documentation types and the function of the dataset (GIS: observations, spreadsheets: financial data, pdf: reports, papers, manuals, html: web pages, etc.).

[2] Due to the limitations of black and white print publication of the proceedings, the colors of the visualization cannot be shown nor explained.



Enter Node Id: 2

Level: 1

W: 1000 H: 750

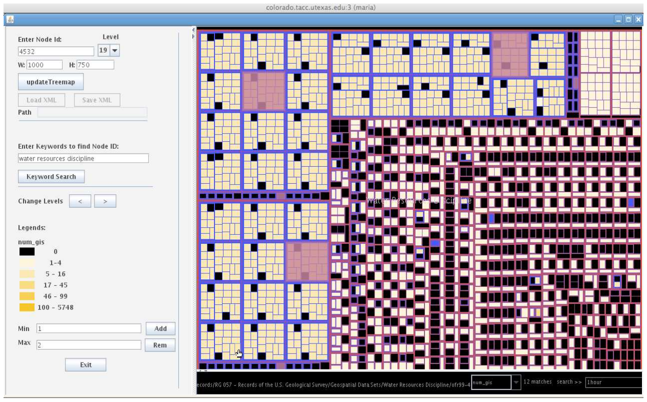
Min: 1

Max: 2

Figure 4. *Distribution of GIS data in the entire collection*

Focused Views

Views can be changed on the fly to show an arbitrary directory and any sub-directory within. Figure 5 below shows one of the directories including subdirectories containing GIS data (upper-left side in figure 4). It is possible to see how the GIS data (highlighted in clear color) is distributed in sub-groups forming distinct patterns corresponding to different datasets or field observations.



colorado.tacc.utexas.edu:3 (maria)

Enter Node Id: 4532

Level: 19

W: 1000 H: 750

Enter Keywords to find Node ID: water resources discipline

Min: 1

Max: 2

ecords/RG 057 - Records of the U.S. Geological Survey/Geospatial Data Sets/Water Resources Discipline/ofr99-4

12 matches search >> 1hour

Figure 5. *Distribution of GIS data in nested directories*

To learn more about the observe patterns we consulted with a GIS data expert who explained that, in order to be effectively processed by GIS software, GIS data has to be arranged in structured fashion. The kinds of arrangement will vary depending on the type and version of the software used. This analysis enables understanding the functions and interoperability needs of a given RG.

Keyword searches

Directory naming is a way of information organization. The visual interface supports keyword searches on the directories' labels at any level of a file path. When a searched term is found, the directory, or group of directories under that term gests highlighted. In figure 5 above and to the left side corner, it is possible to distinguish three sub-groups including one shaded area each. These groupings are named with the same time stamp (1 hour). In turn, the rest of the groups are labeled with different time stamps indicating that this dataset has been consistently organized across the three main groupings.

These focused views of the structure and technical content of groupings of records, from which patterns start to emerge, allow understanding the arrangement of very nested and varied directories as well as the differences between datasets that contain similar data types.

Structure and Patterns

The possibility to represent an heterogeneous collection allows comparing and contrasting different groupings to identify patterns and outliers. Figure 6 below is a close-up of a section of the collection that shows four directories (corresponding to 3 different RGs) with a similar pattern and all of which contain a majority of .pdf files (highlighted in clear color). Beyond sharing the same type of file format, the visual order of the pattern suggests that each dataset is systematically organized.

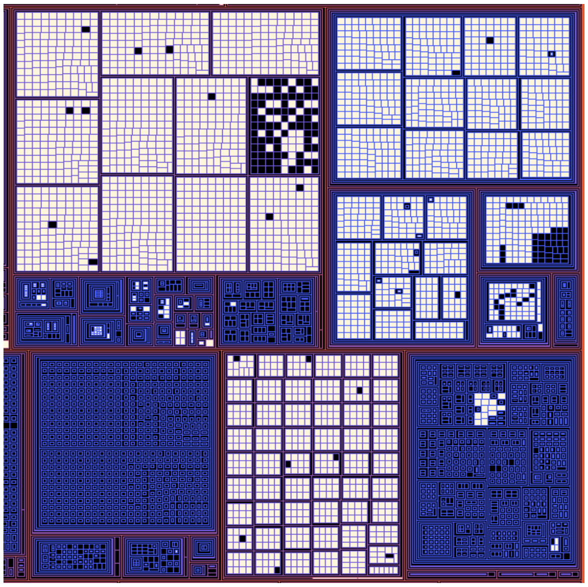


Figure 6. *Identifying patterns and outliers; order from disorder*

To confirm this perception of regular order we browsed across the identified directories with the tooltip and conducted keyword searches to determine the labeling under which the directories are grouped. We found that each of the similar directories is arranged by date and by the type of documentation included: manuals, publications, and statements. In contrast, some surrounding groupings show different file types and irregular patterns.

The intention here was to explore whether by comparing records groupings with similar file formats and visual patterns we can infer which collections belong to a certain "category" of arrangement. The possibility to visually identify with precision—much in the way of a diagnosis,—organized from less organized groupings would require conducting more comparisons between bigger sets of collections. And yet, the possibility to identify structure and to tie that with existing descriptive information helps determining the time and resources needed to give access to them.

Preservation

Preservation decision-making depends first on understanding the collection's condition. We explored the use of information visualization to estimate the overall preservation condition of the collection as well as the distribution of risk in the different RGs considering: a) that directories contain a mix of file formats in different quantities, and b) that the information that we have about the collection is incomplete. In our database we assigned the format score to every object that has one. However, not every object in the collection has been properly identified and not every identified object has a format score.

We generate a view of the entire collection (top) in which users can identify (through color ranges) the percentage of files that have format score at any level of aggregation. In turn, on the bottom view, users can identify the level of risk of a directory, calculated as the average of all the available format scores of the files in that directory. In this way, preservation conditions are derived from those digital objects for which there are reliable information. Figure 7 below shows both presentations for comparison and interpretation purposes.

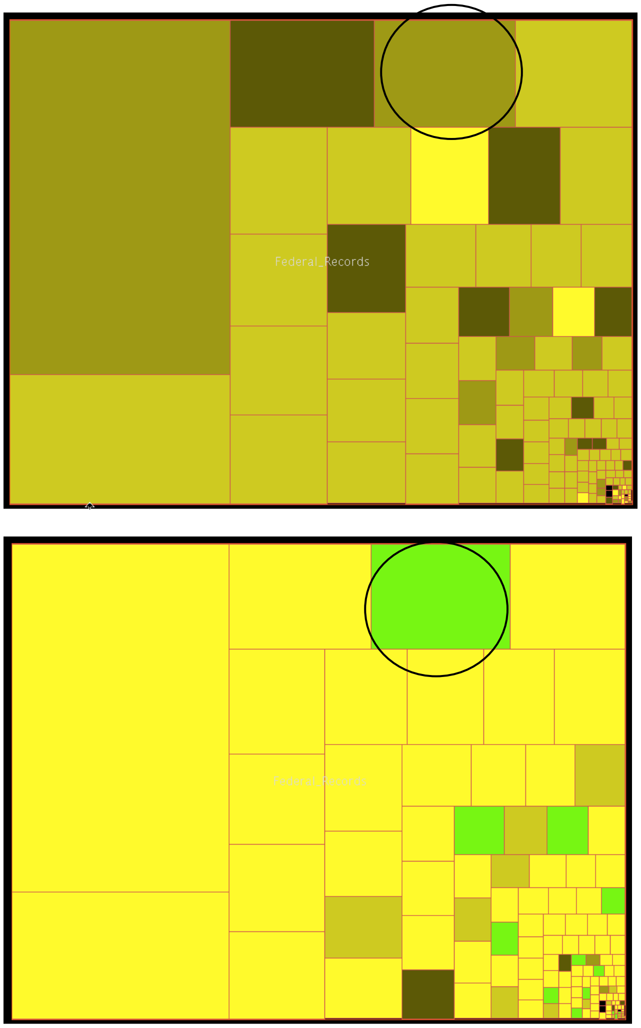


Figure 6. *Preservation sustainability assessment*

A circle in the top points a RG for which 63% of the files have been identified (darker shade). As for the rest, the clearer the shade, the more information we have about the sustainability of their file formats. In the bottom view, the brighter shading indicates that all the files in that RG have high preservation quality. Both views are complementary, as we have to consider that we don't have information for 40% of the files in that RG.

Another way of looking at the information is to observe the distributions of risk and of file category type in the RG. Figure 7 below shows the distribution of risk (top), and the distribution of the predominant file category (bottom) in each directory of the RG.

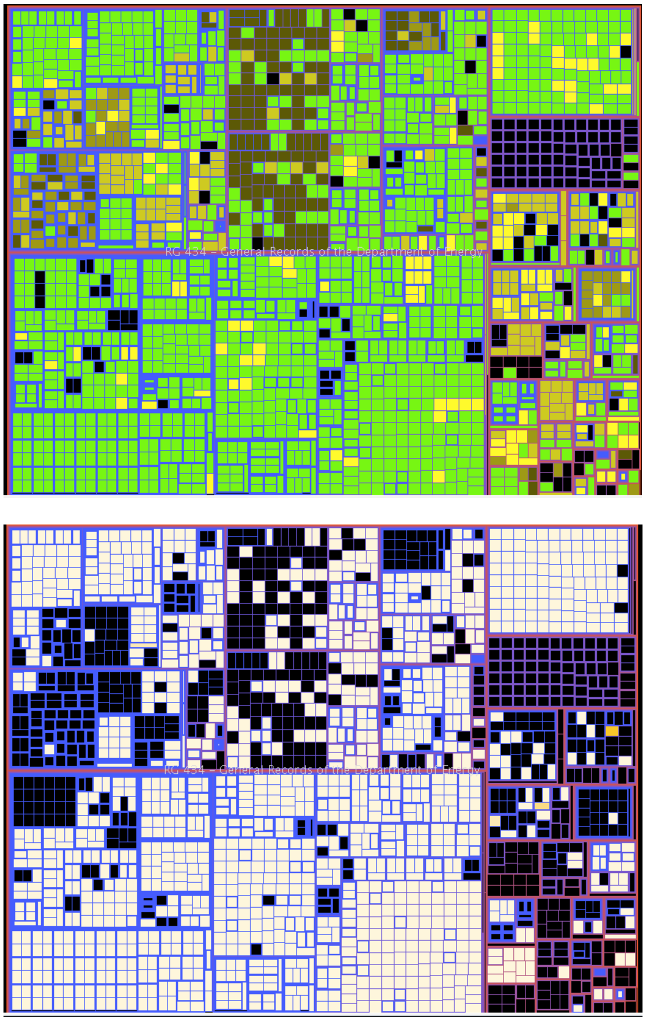


Figure 7. *Distribution of risk (top) and of web file types in a RG.*

In the top view it can be observed that the majority of the directories contain files of high preservation value. In this distribution it is also possible to see the areas for which there is not at all or not much file format score information. In the view at the bottom, the image shows that the web category files (clear shade) are predominant. This category includes html, xml, and css, all file formats of high preservation values.

Appraisal

Presently, by interacting with the visualization features, a user can:

- Evaluate the scope of the entire collection and its sections in terms of number and sizes of files.
- Determine the scope of the different RGs by comparison.
- Study the structure/arrangement of the directories including the configurations of the sub-groups within.
- Determine broad categories of file formats across directories.
- Identify file-naming conventions for directories at the different hierarchical levels considering that file naming constitutes a way of grouping and description.
- Identify similar subjects across the collection according to directory naming conventions.
- Compare and contrast the arrangement of different directories and establish needs for description and access.
- Categorize types of arrangements.
- Associate arrangement with file types and themes across directories and infer the functions of the data types.
- Identify preservation sustainability.
- Compare the preservation sustainability of the different directories and establish preservation priorities.
- Identify distribution of risk within and across directories.
- Identify what is not known about the collection in terms of file formats and preservation sustainability.

The extracted metadata, managed by the RDBMS provides many possibilities to generate different and complementary views of the collection. Combining these views enable to perform various kinds of analysis, each of which provides a piece of information needed to assess the long-term retention of a collection. These analyses are the starting point for archivists to complement with information from additional sources and their own expertise. As opposed to making appraisal decisions based on sampling, the information visualization allows making decisions based on statistical summaries that clearly point to what is known and what is not known about a collection.

Conclusion

The challenges presented by the deluge of digital collections need to be addressed with data driven tools and methods that take advantage of the size and diversity of the data to learn from it. At the same time, data has to be presented in ways that facilitate understanding and enrich the analysis. Our approach is especially fitting to show general trends in the data as well as to detect anomalies. The visual representation generates summarized views of directories, and the values used to determine the rendering are aggregated based on the entire data structure without the need of sampling. Furthermore, the visualization approach also enables users to control the levels of detail to be shown and to make comparisons among data objects. This is important for purposes of detecting outliers and less likely to be achieved by sampling based methods.

Acknowledgments

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.

References

[1] Richard, J., Cox. Appraising the Digital Past and Future. DigCCur 2007. Proc. of DigCCur 2007, an International Symposium on Digital Curation, Chapel Hill, North Carolina. School of Information and Library Science, University of North Carolina at Chapel Hill, (2007), http://www.ils.unc.edu/digccurr2007/index.html (April 18, 2010).

[2] Appraisal Task Force. A Model of the Selection Function, InterPares 1 Project, (2001), http://www.interpares.org/ip1/ip1_aptf.cfm (April 18, 2010).

[3] Meehan, J. Making the Leap from Parts to Whole: Evidence and Inference in Archival Arrangement and Description, The American Archivists, Vol. 72 (Spring/Summer 2009): 72-90.

[4] J, Thomas, and Cook K, eds. Illuminating the Path: The R&D Agenda for Visual Analytics. Research and Development Agenda for Visual Analytics. National Visualization and Analytics Center, (2005). http://nvac.pnl.gov/agenda.stm (April 18, 2010).

[5] The National Archives Center for Advanced Systems and Technologies, NCAST Advanced Research, http://www.archives.gov/ncast/advanced-research.html (April 18, 2010).

[6] IRODS: Data Grids, Digital Libraries, Persistent Archives, and Real-time Data Systems. iRODS. https://www.irods.org/index.php (April 18, 2010).

[7] DROID. Computer software. Vers. 4.0. The National Archives of the United Kingdom, PRONOM, (2006). http://droid.sourceforge.net/ (April 18, 2010).

[8] Anderson, R. et al The AIHT at Stanford University. Automated Preservation Assessment of Heterogeneous Digital Collections, D-Lib Magazine, (2005). http://www.dlib.org/dlib/december05/johnson/12johnson.html, (April 18, 2010).

[9] NDIIP, Sustainability of Digital Formats. Planning for Library of Congress Collections. http://www.digitalpreservation.gov/formats/ , (April 18, 2010).

[10] Shneiderman, B. Treemaps for Space-Constrained Visualization of Hierarchies. A History of Treemap Research at the University of Maryland. http://www.cs.umd.edu/hcil/treemap-history, (April 18, 2010).

[11] Heer, J., Card, S. PREFUSE: A Toolkit for Interactive Information Visualization. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp.421-430, Portland, Oregon, USA: ACM, (2005).

Authors Biographies

Weijia Xu received his PhD in Computer Sciences from the University of Texas at Austin. He is a Research Associate at TACC. Dr. Xu has published in efficient proximity search methods for information retrieval, scientific database management and information visualization.

Maria Esteva received her PhD in Information Science (2008) from the University of Texas at Austin. Since then she works at TACC where she is member of the Data and Collections Management and the Data Analysis and Visualization Groups.

Suyog Jain Dott is a Master's Candidate in the Department of Computer Sciences at the University of Texas at Austin. He is a Graduate Research Assistant at TACC.