InterPARES 3 Project

International Research on Permanent Authentic Records in Electronic Systems

TEAM Canada

Title: General Study 19 – File Viewers: General Study Report

Status:	Final (public)
Version:	1.2
Dated Submitted:	February 2012
Last Revised:	February 2016
Author:	The InterPARES 3 Project, TEAM Canada
Writer(s):	Lois Enns Records Manager, City of Surrey Gurp Badesha Functional Application Specialist, City of Surrey
Project Unit:	Research
URL:	http://www.interpares.org/display_file.cfm?doc=ip3_canada_gs19_final_report.pdf

Document Control

Version history			
Version	Date	By	Version notes
1.0	2012-02-24	L. Enns & G. Badesha	Draft final report.
1.1	2012-02-24	A. Allen	Minor formatting changes.
1.2	2016-02-16	R. Preston	Minor content and copy edits for public version.

Table of Contents

General Study Report

A. Background and Rationale

During the InterPARES 3 TEAM Canada case study on preparing shared drive files for migration into an electronic content management (ECM) system (Rogers *et al.*, 2010), the co-investigators at the City of Surrey identified and adopted a number of utility applications to expedite their work. These utility applications included: a disk space manager, used to collect drive statistics, analyze file formats, create historical profiles, and facilitate metadata discovery; a file manager, used to apply a unique identifier and rename records; a duplication finder, used to identify and remove duplicates based on byte-by-byte comparison; a format identifier, used to identify and resolve missing file extensions; and a empty folder identifier, used to count and (initially) remove empty folders. The use of these utility applications is described in the *Shared Drive Migration Toolkit* (Enns and Badesha, 2011). During the course of the migration work, over 285 file formats were identified and appraised, however, following the business, technical, and records analysis described in the toolkit, only 47 file formats were confirmed as records suitable for migration, and only two of these file formats were found to be obsolete (unpublished data). These two file formats (.ptn and .dwt) represented only 18 files out of 98,000 appraised for migration. The balance of the file formats confirmed as records could be opened, dependent on the availability of the native application.

During the shared drive migration and ECM implementation, the City of Surrey team made a decision not to convert any files to long-term preservation formats. Beyond the constraints of time and available resources, the shared drive environment was ill-suited to linking the native and preservation files, and managing the associated metadata. By contrast, specific features in the ECM system could be leveraged to meet these ends post-migration. Additionally, in an active record system, there were difficulties in identifying the point-in-time at which a file should be put on a migration pathway. Although some record series could be converted to a preservation format on a time-based schedule, other record series reflected event-based activities where secondary files suitable for migration were generated for other purposes during the record lifecycle and could be co-opted for preservation. Finally, given the small number of formats identified as being patently obsolete (i.e., two), there did not appear to be a pressing need to begin bulk conversions at this time. In fact, the co-investigators wondered whether the question

of file conversion might be postponed indefinitely, given the ability to open the native files. In relation to this, the ECM system purchased by the City of Surrey included a file viewer that allowed users to open and annotate specialty drawing files where they did not have the native application loaded on their computer. Although subsequent testing revealed that the viewer module was not well integrated to the ECM system and it was not adopted, the idea that a file viewer might extend the life of a file format was appealing. As a secondary consideration in favour of investigating file viewers, the co-investigators found that during pre-migration file renaming activities, opening files to validate contents was a time-consuming activity, since only a few applications could be kept open on the task bar, and time was spent waiting for applications to open or load, and in flipping between native and utility applications. A file viewer that would enable viewing of multiple formats from a single point appeared to be an avenue worth pursuing.

Seeking to investigate file viewer further, the co-investigators worked with InterPARES 3 to formulate research questions, and four areas of interest were identified: how do file viewers work; what software is available for use; how accurately do file viewers render files; and what role might file viewers play in digital preservation. Over the course of a year, these questions were examined by the co-investigators and discussed at InterPARES 3 bi-annual workshops. From the outset, the research questions represented an unexplored area of interest, since little was available on the topic, and none of the case study participants or researchers used file viewers in their work at that time, with the exception of a research assistant who had used a file viewer while working at the Washington State Archives.

Literature Review

A number of articles mentioning file viewers are found in software and computer engineering journals, primarily with respect to the role of file viewers in software design. For example, an article on a product called GroupKit mentions a file viewer in the context of enabling users' views of text documents in a Unix conferencing environment (Roseman and Greenberg 1995, p.6). Similarly, two articles mention file viewers in the context of Unix programming, and list other types of viewers: a directory viewer, an error viewer, an execution viewer, a software landscape viewer, and an interface viewer (Manoridis *et al.*, 1993, pp.16, 18) and a project viewer and a graph viewer (Anderson and Teitelbaum, 2001, p.3). Interestingly, file viewers are one of a

number of viewers used to interpret machine language into human-readable form.

Adjacent to this work are articles on file format identification (later found to be a component of file viewing). There are at least three computer-based methods for determining file formats: extension-based detection; magic-numbers-based detection; and content-based detection (Amirani *et al.*, 2008). Essentially, the extension-based approach uses file names and mime types; the magic number approach uses the "secret" numbers hidden in file headers; and the content- based approach references "fileprints" through different types of frequency analysis (McDaniel and Heydari, 2002 and Amirani *et al.*, 2008). Scattered through these highly technical articles are suggestions as to why file format identification work is important, including: detection of changes made by a malicious user; dealing with proprietary file types; obsolescence (Dhanalakshmi and Chellappan, 2009); and the need "to preserve data beyond the life of a particular piece of software" (McHenry *et al.*, 2009).

Within the format-identification articles, "Towards a Universal, Quantifiable, and Scalable File Format Converter" (McHenry *et al.*, 2009) is of particular interest. Here, the authors express concern that since "not every format supports the same data content" (p.140), data is dropped when a file is converted from one format to another. In order to minimize the data lost during conversions, they propose an NSCA Polyglot, or "a framework for measuring the quality of individual conversions and allowing for the use of this information in choosing optimal conversion paths" (p.146). They note that, "Aside from the ability to convert between many formats another useful application of such a potentially 'universal' converter is in the form of a 'universal viewer.' Given the ability to view one format in each domain, one could potentially view them all with such a converter by converting every file to this target format…" (p.146). With many archival and records institutions following a migration pathway strategy to long-term digital preservation, a universal file viewer that converts source formats to destination formats "on the fly" presents intriguing new possibilities.

Focusing on file formats, a number of articles and project reports in the library and archives realm examine the significant properties of file formats or "the characteristics of digital objects that must be preserved over time in order to ensure the continued accessibility, usability, and meaning of the objects" (Wilson, 2007a, p.15). Many digital preservation projects (e.g., Investigating Significant Properties of Electronic Content over time, Creative Archiving at Michigan and Leeds Emulating the Old On the New, CURL Exemplars in Digital Archives,

Preservation and Long-term Access through Networked Services) and national archives (e.g., National Archives of Australia, National Archives and Records Administration, The National Archives) have published papers or web articles on significant properties (also called "significant characteristics" or "essential characteristics") in the context of providing a means of measuring whether a preservation strategy such as migration or emulation is successful, by comparing how well a target file retains the properties found in the source file. The "Significant Properties Report" (Wilson, 2007b) provides a useful overview, beginning with a reference to "Canonicalization: A Fundamental Tool to Facilitate Preservation and Management of Digital Information" in which Lynch notes, "We want to be able to guarantee that for a given object the reformatted version is equivalent to the original version with regard to some specific set of object characteristics" (as quoted in Wilson, 2007b, p.5).

An important shift in the significant properties discussion came with the general acceptance that digital objects "do not need to remain in a state that is unchanged from their original state in order for them to be considered authentic" (Wilson, 2007b, p.4). Instead, "A record is considered essentially complete and uncorrupted if the message meant to communicate in order to achieve its purpose is unaltered" (The National Archives, 2002, p.8, as quoted in Wilson, 2007b, p.4). However, an ensuing problem results, because what is considered "essential" may vary from audience to audience. For example, when looking at medieval manuscripts, an audience interested in text analysis would consider the text of a document to be essential, while an audience interested in literary metaphor would insist that the illustrative and design components as important as the text. Unfortunately, despite a "pressing need" to "develop a methodology, and begin identifying quantifiable sets of significant properties for specific classes of digital object[s]" (Wilson, 2007b, p.7), there is no definitive set of significant properties available.

Although some studies provide examples of significant properties for audio, email, raster images, and structured text (Grace, 2009), the *InSPECT Framework Report* reflects a general move towards developing a methodology or framework whereby "an evaluator operating in a curatorial institution can determine the properties that they consider to be essential based on their interpretation of acceptable loss" (Knight, 2009, p.9). To this end, institutions such as the Library of Congress and the Florida Digital Archives have identified and posted the significant properties of interest to their institution on their websites.

B. Methods

Essentially, the research project relied selecting a workable number of file viewer products, formats and files, and comparing each file displayed in the native application with the same file displayed in a range of file -viewer software products. There were five phases to the project: product selection; software comprehension; test environment set-up; file formats, file properties, and files selection; and testing and data collection. With the exception of software comprehension, all phases were run twice, with adjustments made between the two runs. At the end of the project, the results from the second run were reviewed to see how well each file viewer performed, and a determination was made as to whether the file viewer "passed" or "failed" with regard to each file format.

Product Selection

The main criteria for selecting the file viewer products included affordability, ease-of-access, and number of format categories covered. These three criteria can be explained as follows. First, in the previous shared drive migration project, none of the five utility applications cost more than $100 US per licence, and based on this experience, this amount was used as the cut-off point for affordability. Second, due to information technology policies at the City of Surrey, unsupported software cannot be loaded to office computers, and software is only considered supported after being put through an internal review processes. Lacking a guarantee that any of the software would be adopted by the co-investigators' organizational unit, a personal laptop was used for file viewer loading, meaning that only software compatible with a Windows environment was loaded, and that higher-level programming abilities were not required for the install. Third, format categories were loosely considered to be text, data, email, drawing, still image, and moving image, and any file viewer selected was required to provide services for at least two categories. Whether or not the file viewer offered integration with the ECM product was noted, but if the file viewer met the other criteria, it was selected, regardless.

With these criteria in mind, a number of Google searches (e.g., "file viewers," "universal viewers," "best file viewers") were completed, resulting in a list of possible products. Next, Download.com and SorceForge.net were used as qualifying resources. Download.com features software reviews, technology news and software downloads. Their section on universal viewers helped the co-investigators short list the top 15 products within the cost requirement. Once the

products were short-listed, each product website was reviewed to identify the best fit for the project, and the final selection was made. SouceForge.net, a site for open-source software development, was also referenced, but although open-source file viewers were identified, only one open-source file viewer indicated that two format categories were supported, and an attempt to download this product was unsuccessful due to programming requirements.

Software Comprehension

Once the file viewers were identified, the product websites were reviewed in an attempt to find information on how the products work. Since these are proprietary products, little was available. Next, the co-investigators contacted the software developers directly, using email and web forums. In every case, the developers were advised that the co-investigators were seeking information for a research paper on file viewers.

Test Environment Set-Up

Two computers were used in the test environment: a Windows-platform workstation connected to the City of Surrey's networked computing environment; and a Windows-platform personal laptop not connected to the network. All of the test files were maintained on the workstation, and all of the file viewers were maintained on the personal laptop. The test files were transferred from the workstation to the personal laptop using a USB drive.

File Formats, File Properties and Characteristics, and Files Selection

Although there are currently over 40 formats available in the ECM system, only the top 12 formats were selected for testing, with each represented by at least 500 files and up to 18 years' worth of files (with one exception). In this respect, the file formats selected reflect an intent to test file viewers, rather than file formats. Perhaps focusing on popular file formats provided the file viewers products with a better probability of success, but this was considered acceptable based on the discovery nature of the project. As well, the choice of popular file formats was dependent on the particular business units whose files were stored in the ECM system, which currently supports about 10 percent of the networked staff population. The selection of file formats chosen at another organization might differ (as would a selection made after more business units are added to the ECM system), but, based on the City of Surrey file format list, the co-investigators are inclined to believe that many of their popular formats would

be found at other facilities.

Lacking a significant properties standard, the co-investigators referenced the significant properties listed on the InSPECT, Florida Digital Archives (FDA), and the Library of Congress (LOC) websites for each format category, as available: text (InSPECT, FDA, LOC); data (FDA); email (InSPECT); web (InSPECT); image (InSPECT, FDA, LOC); and moving image

(InSPECT, LOC). Based on their experience in the previous shared drive migration project, the co-investigators divided significant properties (as used by InSPECT, FDA, and LOC) into *properties,* which could be determined without opening a file, and *characteristics,* which could only be determined by opening a file. For all format categories, three properties were identified: Title, Creator, and Date Created. Additional properties were identified by format category: Word Count (text); Resolution, Bit Depth, Width, and Height (image); and Length, Width, Height, Pixel Aspect Ratio, and Frame Rate (moving image). In each case, these properties could be viewed using the Windows operating system and/or a file manager utility application, and the native application. What this means is that the co-investigators' analysis of file viewer performance was not be based on whether or not metadata properties could be accessed, but rather whether or not characteristics were observed.

These properties did not reveal much about the ability of a file viewer to render files. Instead, with reference to the InSPECT, FDA, and LOC websites, and the co-investigators own observations, a list of characteristics was designed to test the file viewers from the inside. These characteristics included: Header and Footer, Font Size and Colour, Images/Diagrams, Bullets and Numbering, Print, *Hyperlinks, Page Count, and Text Search* (text); *Font Size and Colour,* Cells, *Formulas, Macros and Links, Frames/Page Breaks* (data); Font Size and Colour, Sender, Receiver, Name, Date Sent, Date Received, Subject, Attachments, Body, Signature (email); Division, Paragraph, Image, Link, Frame (web); Font Size and Colour; Colour, Scalability, Sharpness, *Page Number* (drawing); Colour, Completeness (image); and Colour, Sound, and Back and Forward Navigation (moving image). (Note that the mandatory characteristics are shown in regular font, while the *optional characteristics* are shown in italic font.)

A first run of testing was completed on seven file viewers and 14 file formats for a total of 126 files. For each format, nine files were selected, with three files selected for each of three time blocks (1994-1999; 2000-2005; and 2006-2011). These time blocks were intended to demonstrate whether file viewers were to any degree backwards compatible. In the second run of

testing, only nine files were identified for each of 12 formats for a total of 108 files. Significant care was taken to ensure that each of the files was an appropriate candidate for testing and presented properties and characteristics of interest.

Data Collection

Data was collected for using a separate chart for each file format and each file viewer, for a total of 72 charts (see Tables 3 to 5). These results were tabulated in pass/fail summary charts, separated into the three time periods (see Tables 6 to 8).

C. Results

Product Selection Results

Based on the three criteria of affordability, ease-of-access, and minimum format categories, six file viewer products were identified and tested in the second run:

1. Accessory Software File Viewer ($23.00);
2. FileStream Turbo Browser ($69.00);
3. GetData Explorer View ($29.95);
4. Irfan View ($10.00 donation);
5. Quick View Plus ($49.00); and
6. UV ViewSoft ($25.00).

Trial versions were available for a number of products, and a number of these versions were loaded and previewed to help the co-investigators familiarize themselves with this file viewers in general. A trial version of a product called Daeja ViewONE was used in the first run testing, but later discarded due to pricing ($1,200 for a site licence).

In reviewing file viewer products, it appeared that there were two categories of file-viewer software: low-cost file viewers intended as stand-alone products; and more costly file viewers intended for integration with other software.

Software Comprehension Results

Altogether, five companies responded to the co-investigators questions regarding how file viewers work: Accessory Software File Viewer, Daeja ViewONE, IrfanView, Oracle Stellant, and UV ViewSoft. There were two barriers to learning how file viewers work: first, companies protect proprietary information; and, second, as one contact noted, “it is tricky to get

the fine detail on this as it is quite a complex subject." However, by piecing together the information provided by the software developers, a basic understanding was achieved.

In general, file viewers work by identifying file formats through header information, magic numbers, or content, and then rendering the content in human-readable form. If the file format is one that the viewer can render "as is," the file is displayed in native format. If not, the file is converted to a second format and then rendered. In order to extend their file format rendering capabilities, file viewers often consist of a number of viewers bundled together. For example, the Accessory Software File Viewer contact noted their product uses viewers from Internet Explorer (for text, html, and Microsoft Object Linking and Embedding or OLE); LEADTOOLS (for images); and Delphi (for data with open database compliancy or ODBC). Similarly, UV ViewSoft leverages the Microsoft Internet Explorer engine (for html); a doc-rtf converter (for text); and Delphi (for data). The Daeja ViewONE contact referred to "third-party libraries," and the Stellant respondent noted the use of "outside-in" libraries which convert "foreign" formats to a generic format which standard viewers can then access. As a result, it appears that most products rely on file filters and conversion, and are, in the words of the Stellant respondent "actually rendering a much smaller number of standard formats" than the 100 to 300 file formats commonly listed in product information. During testing, it was interesting to note that all six products launched Adobe Reader to render .pdf files.

In some cases, the file viewer product is intended for specific format categories—for example, IrfanView is intended for use with image and audio/video file formats, while FileStream Turbo Browser extends to six format categories. (As a side note, software products that render audio/video file formats are referred to as "players" rather than viewers.) The capabilities noted in product information are summarized in Table 1.

Table 1: File Viewer Products and File Format Capabilities (based on product information)

Products	Text	Data	Email	Drawings	Images	Moving Images
Accessory Software File Viewer	Yes	Yes	No	No	Yes	Yes
FileStream Turbo Browser	Yes	Yes	Yes	Yes	Yes	Yes
GetData Explorer View	Yes	No	Yes	Yes	Yes	Yes
IrfanView	No	No	No	No	Yes	Yes

Quick View Plus	Yes	Yes	Yes	Yes	Yes	No
UV ViewSoft	Yes	No	No	No	Yes	Yes

Some file viewers include additional functionality, such as format conversion, file annotation, file redaction, and integration with electronic record or content management systems. It is assumed that the software developers producing the file viewers look to appeal to a wider market, first by extending the number of file formats that their file viewers can render, and, second, by broadening the marketability of the product by adding related features which users may be interested in. These features included format conversion, editing, annotation, redaction, and product integration, and were more commonly available in the category file viewers that were more costly and likely intended for integration with other software. As shown in Table 2, few of these features were available in the file viewers selected for testing in this project.

Table 2: File Viewer Products and Additional Features (based on product information)

Products	Format Conversion	Edit	Annotation	Redaction	Product Integration
Accessory Software File Viewer	No	No	No	No	No
FileStream Turbo Browser	Yes	Yes	No	No	No
GetData Explorer View	No	No	No	No	No
IrfanView	No	No	No	No	Yes
Quick View Plus	No	No	No	No	No
UV ViewSoft	No	No	No	No	Yes

Based on the product information available, the co-investigators expected that FileStream Turbo Browser and Quick View Plus to perform the best during the data collection phase of the project.

Test Environment Set-Up Results

In general, the test environment, consisting of a City of Surrey networked workstation and a personal laptop was sufficient. In a more ideal situation, the computers used would be

identical, sharing the same software load and common access to the shared drive and files. During testing, differences in screen size and colour space were manually compensated for, and did not present a barrier to data collection. However, because the personal laptop did not have access to the shared drives, files had to be copied, resulting in changes to critical metadata (i.e. Owner, Date Created, Location). This would not have been the case if both computers were networked workstations.

Mitigating the problem of changed metadata by confirming its availability and values reinforced the co-investigators' belief that metadata properties are different from file characteristics. From their point-of-view, a file viewer does not need to be able to access properties that can be better assessed with a more suitable tool such as an operating system or a file manager utility application.

File Formats, File Properties and Characteristics, and Files Selection Results

The twelve file formats chosen for the second run of testing included: text (.doc, .pdf, .ppt), data (.xls), email (.msg), web (.htm), drawing (.dwg, .vsd), image (.jpg, .tif), and moving images (.mov, .avi). All of these formats met the requirement of at least 500 files and at least three files from each of the three time blocks (with the exception of the .avi format, where only files from the last time block were found).

During preliminary testing, a discovery was made that four out of the six viewers could not render Microsoft files in ".x" file formats (i.e., .docx, .pptx, .xlsx), designed to meet the Office Open XML standard. Additionally, the file viewers that could open .htm files rendered the files as text representations with style tags, without graphic representation. The reason for the "xml" gap in the file viewers is not known. Perhaps the developers of these products do not consider xml file formats problematic, assuming that these files will be viewed using a web browser or editor. Or perhaps the xml file formats are too new, and the development work is not complete. At any rate, the .x file formats were removed from the test sample.

Regarding properties (determined without opening a file), there were few unexpected results. Word Count was not available for view for .pdf files in either Windows or file manager. The same issue appeared for Length, Width, Height, Pixel Aspect Ratio, and Frame Rate for .mov files.

Regarding characteristics (determined by opening a file), there were a number of general

observations. First, where a file viewer could render a file on screen, the file could be successfully sent to Print. Second, the web file format (.htm) and image file formats (.jpg, .tiff) were rendered by all six file viewers, although it should be noted that graphic representation of the web file format (.htm) was not a mandatory requirement. Third, some characteristics, such as slide presentation and animation (.ppt) or formulas, macros, and links (.xls) were not represented by any of the file viewers. Because these characteristics appeared to present a common problem to all viewers, they were treated as non-mandatory. Perhaps these characteristics are proprietary features of the native applications, and require execution rather than simple rendering. Fourth, in most cases, if a file viewer could render a file format, it could open older versions of the file format. Exceptions included GetData Explorer View (for .ppt in 2000-2005 and 1994-1999 and for .xls in 1994-1999) and UV ViewSoft Viewer (for .doc in 1994-1999).

During testing, four .docx files were identified and replaced from the test pool after five file viewers failed to render the files. An additional two files were identified as corrupt as they could not be opened in the native application. These files were removed and replaced.

Data Results

Files were selected on the basis that mandatory properties were present and viewable in the operating environment and/or the file manager utility application and the native application. To ensure that good sample files were selected for the test pool, the data was extracted and collected in summary tables for each file format (see Table 3). In some files, the co-investigators found that mandatory properties were missing or overwritten. In these cases, the files were discarded. A total of 72 properties tables were created (i.e. six file viewers x 12 formats).

Table 3: Operating System Results for .doc Properties (mandatory and *optional*)

DOC		Title	Creator	Date Created	Word Count
1994 to 1999					
	File 1	Tracer Introduction and Configuration.doc	Administrators	1996-08-06 9:00	*206*
	File 2	Instructions to Upgrading Firewall.doc	Administrators	1996-10-03 8:52	*697*
	File 3	DCT CSDC Documentation Amanda 3.doc	SURREY\LSA	1996-08-26 10:49	*5472*
2000 to 2005					
	File 1	DCT Audit Report Procure Audit Report.doc	SURREY\NAJ	2000-01-13 16:21	*4293*
	File 2	Steps for Renaming Production databases.doc	SURREY\BL8	2002-07-09 14:12	*2710*

	File 3	DCT Old Pre 7 4 Documents Cognos 1.doc	SURREY\IAM	2000-01-17 07:10	*363*
2006 to 2011					
	File 1	DCT IP3 Creator Preserver Responsibilities V 03 0.doc	SURREY\LE2	2009-05-22 13:03	*3188*
	File 2	SOW Storage Solution Facilities Plans 2008 08 25 v01 0.doc	SURREY\LE2	2008-08-26 07:27	*935*
	File 3	DCT Master List 2011.doc	SURREY\EAG	2010-12-02 11:00	*3051*

Once the file properties were checked, each format was tested on each file viewer by opening each file in the native application on the networked workstation, and then opening a copy of the file on the personal laptop. The characteristics of the files were compared, and the results were recorded in 72 separate tables (see Table 4).

Table 4: File Viewer Showing Pass Results for Characteristics for .doc File Format

DOC	**Header/ Footer**	**Font**	**Images/ Diagrams**	**Bullets**	***Hyperlink***	***Page Count***	***Text Search***	**Print**	**OCR**
Accessory Software File Viewer									
1994 to 1999									
File 1	PASS	PASS	PASS	PASS	*N/A*	*PASS*	*PASS*	PASS	
File 2	PASS	PASS	N/A	PASS	*PASS*	*PASS*	*PASS*	PASS	
File 3	PASS	PASS	PASS	PASS	*PASS*	*PASS*	*PASS*	PASS	
2000 to 2005									
File 1	PASS	PASS	PASS	PASS	*PASS*	*PASS*	*PASS*	PASS	
File 2	PASS	PASS	PASS	PASS	*PASS*	*PASS*	*PASS*	PASS	
File 3	PASS	PASS	PASS	PASS	*N/A*	*PASS*	*PASS*	PASS	
2006 to 2011									
File 1	PASS	PASS	PASS	PASS	*PASS*	*PASS*	*PASS*	PASS	
File 2	PASS	PASS	PASS	PASS	*N/A*	*PASS*	*PASS*	PASS	
File 3	PASS	PASS	N/A	PASS	*PASS*	*PASS*	*PASS*	PASS	

The co-investigators determined whether a file viewer passed or failed based on mandatory characteristics. In the comparison test, the mandatory characteristics (shown in regular font) had to be seen in order for the file viewer to receive a pass for that format (see Table 4). If the mandatory characteristics were not present, the file viewer received a fail (see Table 5).

Table 5: File Viewer Showing Fail Results for Characteristics for .doc File Format

DOC	Header/ Footer	Font	Images/ Diagrams	Bullets	*Hyperlink*	*Page Count*	*Text Search*	Print	OCR
GetData Explorer View									
1994 to 1999									
File 1	FAIL	FAIL	FAIL	FAIL	*N/A*	*FAIL*	*PASS*	PASS	
File 2	FAIL	FAIL	N/A	FAIL	*FAIL*	*FAIL*	*PASS*	PASS	
File 3	FAIL	FAIL	FAIL	FAIL	*FAIL*	*FAIL*	*PASS*	PASS	
2000 to 2005									
File 1	FAIL	PASS	PASS	FAIL	*FAIL*	*FAIL*	*PASS*	PASS	
File 2	FAIL	PASS	PASS	FAIL	*FAIL*	*FAIL*	*PASS*	PASS	
File 3	FAIL	PASS	PASS	FAIL	*N/A*	*FAIL*	*PASS*	PASS	
2006 to 2011									
File 1	FAIL	PASS	PASS	FAIL	*FAIL*	*FAIL*	*PASS*	PASS	
File 2	FAIL	PASS	PASS	FAIL	*N/A*	*FAIL*	*PASS*	PASS	
File 3	FAIL	PASS	N/A	FAIL	*FAIL*	*FAIL*	*PASS*	PASS	

Once the data for characteristics were collected, the pass/fails for each file viewer and file format were compiled into three charts, showing the performance of the file viewer on newer files dated from 2006 to 2011, somewhat older files from 2000 to 2005, and older files from 1994 to 1999 (see Tables 6-8).

There were no file viewers that could successfully render all 12 file formats. Two file viewers were able to open 10 out of 12 formats: FileStream Turbo Browser and Quick View Plus. Turbo Browser was unable to open .doc and .vsd files, Quick View was unable to open .mov or .avi files. Of these two products, Quick View performed more closely to product claims as interpreted by the co-investigators (see Table 1).

Table 6: File Viewer Capabilities by File Format (2006-2011)

File Viewer	DOC	PDF	PPT	XLS	MSG	HTM	DWG	VSD	JPG	TIFF	MOV	AVI
Accessory Software File Viewer	PASS	PASS	FAIL	PASS	FAIL	PASS	FAIL	FAIL	PASS	PASS	FAIL	PASS
FileStream Turbo Browser	FAIL	PASS	PASS	PASS	PASS	PASS	PASS	FAIL	PASS	PASS	PASS	PASS
GetData Explorer View	FAIL	PASS	PASS	PASS	FAIL	PASS	FAIL	FAIL	PASS	PASS	FAIL	FAIL
Irfan View	FAIL	FAIL	FAIL	FAIL	FAIL	PASS	FAIL	FAIL	PASS	PASS	FAIL	PASS
Quick View Plus	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	FAIL	FAIL
UV Viewsoft Viewer	PASS	PASS	FAIL	PASS	FAIL	PASS	FAIL	FAIL	PASS	PASS	PASS	PASS

Table 7: File Viewer Capabilities by File Format (2000-2005)

File Viewer	DOC	PDF	PPT	XLS	MSG	HTM	DWG	VSD	JPG	TIFF	MOV	AVI
Accessory Software File Viewer	PASS	PASS	FAIL	PASS	FAIL	PASS	FAIL	FAIL	PASS	PASS	FAIL	N/A
FileStream Turbo Browser	FAIL	PASS	PASS	PASS	PASS	PASS	PASS	FAIL	PASS	PASS	PASS	N/A
GetData Explorer View	FAIL	PASS	FAIL	PASS	FAIL	PASS	FAIL	FAIL	PASS	PASS	FAIL	N/A
Irfan View	FAIL	FAIL	FAIL	FAIL	FAIL	PASS	FAIL	FAIL	PASS	PASS	FAIL	N/A
Quick View Plus	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	FAIL	N/A
UV Viewsoft Viewer	PASS	PASS	FAIL	PASS	FAIL	PASS	FAIL	FAIL	PASS	PASS	PASS	N/A

Table 8: File Viewer Capabilities by File Format (1994-1999)

File Viewer	DOC	PDF	PPT	XLS	MSG	HTM	DWG	VSD	JPG	TIFF	MOV	AVI
Accessory Software File Viewer	PASS	PASS	FAIL	PASS	FAIL	PASS	FAIL	FAIL	PASS	PASS	FAIL	N/A
FileStream Turbo Browser	FAIL	PASS	PASS	PASS	PASS	PASS	PASS	FAIL	PASS	PASS	PASS	N/A
GetData Explorer View	FAIL	PASS	FAIL	FAIL	FAIL	PASS	FAIL	FAIL	PASS	PASS	FAIL	N/A
Irfan View	FAIL	FAIL	FAIL	FAIL	FAIL	PASS	FAIL	FAIL	PASS	PASS	FAIL	N/A
Quick View Plus	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	FAIL	N/A
UV Viewsoft Viewer	FAIL	PASS	FAIL	PASS	FAIL	PASS	FAIL	FAIL	PASS	PASS	PASS	N/A

D. Conclusion

The project provided relatively straight-forward results for the first three research questions. First, file viewers were found to work by rendering a file in its native file format, or by converting a file to a second format that could be rendered. Second, file-viewer software products appear to fall into two categories: lower-cost, stand-alone products; and higher-cost products capable of integration with electronic document, record, or content management systems. In general, the six lower-cost, stand-alone products tested performed as per the product claims. Differences between product claims and performance may depend on the metrics chosen

as there is no standard set of properties and characteristics against which performance is consistently measured. Third, with respect to accuracy, every file viewer provided 100 percent accuracy for at least two of the seven file format categories (text, data, email, web, drawing, image, moving image), for the 12 file formats tested, based on the properties and characteristics identified by the co-investigators. Specifically, Quick View Plus provided 100 percent accuracy for six format categories over three time blocks; FileStream Turbo Browser, for five categories; UV Viewsoft Viewer, for four categories; and Accessory Software File viewer, GetData Explorer View, and Irfan View for two categories. Importantly, the file viewers performed better than expected in terms of rendering older files, with only two file viewers showing reduced performance over three time blocks.

Regarding the question as to whether file viewers might play a role in digital preservation, the co-investigators can offer some thoughts, based on their experiences providing record management services to local government staff. Digital preservation aside, the co-investigators will find an immediate use for the two products identified within their shared drive appraisal work, including two projects consisting of 29,000 and 20,000 files. There is also a possibility that one or both of the products will be of interest to other organizational units for access to files where software licensing costs represent a barrier to read-only use. (A good example is the .dwg file format, created by architectural drawing programs such as AutoCAD and widely used in the Building, Planning, Civic Facilities, Engineering, and Realty Services divisions.)

The co-investigators do see file viewers as offering an under-examined opportunity for a new or complimentary digital preservation strategy. Currently, the migration pathway strategy, where native file formats are converted to preservation formats (and, in the OAIS model, dissemination formats) is not an ideal fit for an active record environment like the City of Surrey. Practical issues include shortages with respect to storage space, technical capabilities, staff time, and work processes. For example, creating preservation copies for all permanent unstructured electronic records require significantly more storage and current requirements may be taxing the existing capacity. The existing software products in place may not provide migration management functionality, and the organization may have a policy against adopting open-source software common in file migration. The information technology unit may not support implementation or provide necessary resources. And, lacking practical examples for

implementing a migration pathway approach in an active record environment, the records management staff may not be able to develop a workable procedure for creating, monitoring, and managing preservation copies. By comparison, a file viewer requires only installation and minimal training, and supports complimentary uses in terms of file format appraisal.

Beyond the logistic problems, migration conversion from a native to a preservation format often involves loss, since "not every format supports the same data content" (McHenry, Kooper, and Bajcsy, p.140). Recognized data loss resulting from .pdf conversion include formulas (.xls to .pdf), slide presentation and animation (.ppt to .pdf), and hyperlinks (.doc to .pdf). Although some researchers do express the belief that digital objects "do not need to remain in a state that is unchanged" (Wilson, 2007b, p.4), over ten years ago, researchers on the CAMiLEON project noted that "Existing methods of preserving digital data often fall short of accurately preserving and authentically rendering an original digital document..." and that "There are many drawbacks with this strategy of 'traditional migration'... Any errors or omissions from a transformation will propagate..." (Mellor *et al.*, 2002, p.517). In the CAMiLEON project, "migration on request" was proposed as an alternative strategy to migration conversion. Here, a "digital object is simply archived in its original format," based on "the principle of always maintaining the original bytestream" (p.518) and the bar was raised to the point where only "the only way of ensuring a migration step has been completed without error is by the proof of reversible migration" (p.519). As unattainable as this may sound, this approach was successfully tested using a custom-built Migration on Request tool, with a focus on Scalable Vector Graphics (SVG) features (p.522-524). Why this is of interest in the file viewer context is that the co-investigators' research showed that, at the current time, a file-viewer digital presentation strategy would present the same problem as a migration-conversion preservation strategy, in that similar data is lost in both scenarios. So in this regard, a file viewer appears to provide similar performance to migration conversions of the same type.

In all cases identified, data loss related to characteristics rather than properties. In the context of properties, using file viewer means that the native file is viewed, and the properties represent the original values (assuming they have been appropriately preserved). By comparison, the preservation file will retain some of the original properties, but overwrite others, including Owner and Date Created. However, this point is, to some extent, irrelevant. While properties are fragile and must be carefully maintained to retain original values, whether or not properties can

be viewed from within a file is, in the minds of the co-investigators, irrelevant. If the properties exist, are unchanged, and can be extracted using the operating system or a file manager, their maintenance is likely a separate issue than whether or not data characteristics are evident in a migration copy. Characteristics need to be rendered to be observed, whether rendered on the fly by a file viewer or created through conversion. Based on this understanding, the co-investigators suggest that a file-viewer approach has a minor advantage over a migration-conversion approach as the original properties are evident as they are embedded in the single file. In migration conversion, the relationship between both the files must be created and maintained, along with both sets of properties.

Returning to characteristics, the file viewers displayed similar issues around data loss as experienced in migration conversion. In this respect, it is interesting to note that where the original developer's rendering software was available, all six file viewer developers adopted it as the rendering "engine" for that format. During testing, the co-investigators found that where a .pdf file was launched using a file viewer, the file opened in Adobe Acrobat Reader. One wonders, if all software developers made their rendering engines available (as Adobe has), whether file-viewer developers might not be able to create products closer to a "universal" viewer, capable of rendering all (or at least more) formats. Additionally, might not some of the characteristics that render (or play) within the native application and that are not apparent in file viewer or migration conversions be made available from within the rendering engines to provide a fuller data range for these objects? After all, the Adobe Acrobat business model of providing the "reader" for free and charging for the "writer" proved successful, even after the .pdf specification was made available. Although some type of planetary alignment might be required to achieve this effect, it is not impossible, given the number of archives, records management, and library specialist now involved in standards boards and other organizations that facilitate access to software development companies.

In this respect, it may be of value for the archives and records management community to continue to clarify their requirements in terms of properties and characteristics. Properties, in the sense of file property metadata, are clearly conveyed through standards, data dictionaries, and many other forums, but additional clarification around characteristics would be useful. Here, the co-investigators observe three categories of characteristics: structure-related (e.g., cells, line breaks, page breaks, tables, and bullets); appearance-related (e.g., font size and colour, images,

and diagrams), and behaviour- related (e.g., formulas, macros, and slide presentation and animations). Similar observations regarding categories of significant properties were noted in the *InSPECT Significant Properties Report* (Wilson, 2007b), including content, context, appearance, structure, and behaviour, but no separation was made between properties and characteristics.

In closing, the co-investigators recognize the migration-conversion approach as the primary digital preservation strategy in place in most archive institutions today. This strategy provides important risk insurance for digital objects, and especially those in danger of immediate obsolescence. For some organizations, the risk of not having electronic information available in an accessible format largely outweighs the total costs of file migration. However, other organizations may find that the migration-conversion strategy is not always a viable option, and they still may need to provide some means of long-term preservation of digital objects. The file viewer approach was reviewed by the co-investigators in the face of specific problems in introducing migration conversion to their environment. Additionally, the migration-conversion strategy is not perfect, as data characteristics are often lost. With many institutions maintaining the native files as their only record or in addition to a preservation copy, an opportunity exists to pursue other, complementary strategies. For these two reasons, the co-investigators suggest that file viewers provide an opportunity to leverage native files in a less resource-intensive manner. The co-investigators note also the extensive body of work in progress beyond the field of archives and records management, and the need to collaborate with other fields, including software development in the pursuit of the file viewer and other digital preservation strategies.

E. Acknowledgements

The co-investigators would like to acknowledge the work of the InterPARES 3 Graduate Research Assistants, Jen Busch and Sergey Kovynev, who participated in launching the *General Study on File Viewers,* and the contributions of Dr. Luciana Duranti, InterPARES Project Director.

F. References

"Accessory Software File Viewer 9," available at http://www.accessoryware.com/FileView.htm (accessed November 2011 - February 2012).

Amirani M., Toorani M. and Shirazi A. (2008), "A New Approach to Content-based File Type Detection," in *Proceedings of the 13th IEEE Symposium on Computers and Communications (ISCC'08), July 2008,* pp. 700-705.

Anderson, P. and Teitelbaum, T., (2001), "Software inspection using CodeSurfer," *WISE'01: Proceedings of the First Workshop on Inspection in Software Engineering, Paris, July, 2001,* pp. 1-9.

"CNET Download.com," available at http://www.download.com (accessed November 2011 - February 2012).

"Daeja ViewONE," available at http://www.daeja.com/products/viewone-pro-overview.asp (accessed November 2011 - February 2012).

Dhanalakshmi, R. and Chellappan, C., (2009), "File Format identification and information extraction," available at http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=5393688 (accessed November 2011 - February 2012).

Enns, L. and Badesha, G. (2011), *Shared Drive Migration Toolkit*, City of Surrey, available at http://www.interpares.org/ip3/ip3_cs14_report.cfm (accessed September 2011 - February 2012).

"FileStream Turbo Browser," available at http://www.filestream.com/turbobrowser/ (accessed November 2011 - February 2012).

The Florida Centre for Library Automation, (2009), "Florida Digital Archives," available at http://fclaweb.fcla.edu/FDA (accessed November 2011 - February 2012).

"GetData Explorer View," available at http://www.explorerview.com/ (accessed November 2011 - February 2012).

Grace, S., Knight, G. and Montague, L., (2009), "Investigating the Significant Properties of Electronic Content over time: Final Report," Centre for e-Research, King's College London, 21 December 2009.

Knight, G. (2007), "InSPECT Investigating Significant Properties of Electronic Content," available at http://www.significantproperties.org.uk/ (accessed November 2011 - February 2012).

Knight, G. (2009), "Investigating the Significant Properties of Electronic Content over time: Framework Report," Centre for e-Research, King's College London, 13 October 2009.

Library of Congress, (2010) "Sustainability of Digital Formats: Planning for Library of Congress Collections," available at http://www.digitalpreservation.gov/formats/ (accessed November 2011 - February 2012).

Mancoridis, S., Holt, R. C. and Penny, D. A., (1993), "A 'curriculum-cycle' environment for teaching programming," *SIGCSE '93 Proceedings of the twenty-fourth SIGCSE Technical Symposium on Computer Science Education*, New York, NY, USA, 1993.

McDaniel, M. and Heydari, M. H., (2002), "Content Based File Type Detection Algorithms," available at http://dl.acm.org/citation.cfm?id=821828 (accessed November 2011 - February 2012).

McHenry, K., Kooper, R. and Bajcsy, P. (2009), "Towards a Universal, Quantifiable, and Scalable File Format Converter," available at http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=5380873 (accessed November 2011 - February 2012).

Mellor, P., Wheatly, P. and Sergeant, D., (2002), "Migration on Request, A Practical Technique for Preservation," in *Research and Advanced Technology for Digital Libraries, 6th European Conference, ECDL 2002, Proceedings, Rome, Italy*, 16-18 September 2002, pp. 516-526.

"Quick View Plus," available at http://www.avantstar.com/metro/home/Products/QuickViewPlusStandardEdition (accessed November 2011 - February 2012).

Rogers, C., Malmas, S., and Enns, L., (2011), "Case Study Final Report: CS 14 City of Surrey Policies, Guidelines and Procedures for a Drive Migration Project as part of an Enterprise Content Management Program," available at http://interpares.org/ip3/display_file.cfm?doc=ip3_canada_cs14_final_report.pdf.

Roseman, M. and Greenberg, S., (1995), "Building Real-time Groupware with GroupKit , A Groupware Toolkit," *ACM Transactions on Computer-Human Interaction (TOCHI)*, Volume 3, Issue 1, pp. 1-30.

Skiljan, I., (2005), "Irfan View," available at http://www.irfanview.ca/ (accessed November 2011 - February 2012).

"Sourceforge," available at http://sourceforge.net/ (accessed November 2011 - February 2012).

"Stellant Outside In," available at http://www.oracle.com/us/technologies/embedded/025613.htm (accessed November 2011 - February 2012).

"Universal Viewer," available at http://www.uvviewsoft.com/ (accessed November 2011 - February 2012).

Wilson, A. (2007a), "Significant Properties of Digital Object," *National Archives of Australia*, JISC Significant Properties Workshop, British Library, April 2008.

Wilson, A. (2007b), "InSPECT: Significant Properties Report," *Arts and Humanities Data Service*, 10 April 2007.