Conference Proceedings

Edited by:
Luciana Duranti and
Elizabeth Shaffer

The Memory of the World in the Digital Age: Digitization and Preservation

An international conference on permanent access to digital documentary heritage

United Nations
Educational, Scientific and
Cultural Organization

Memory of the World
20th Anniversary

Hosted by:

a place of mind
THE UNIVERSITY OF BRITISH COLUMBIA

In collaboration with

UNIVERSITY OF TORONTO

26 to 28 SEPTEMBER 2012
Vancouver, British Columbia, Canada
Sheraton Vancouver Wall Centre

Conference Proceedings

Edited by:
Luciana Duranti and
Elizabeth Shaffer

The Memory of the World in the Digital Age: Digitization and Preservation

An international conference on permanent access to digital documentary heritage

26 to 28 SEPTEMBER 2012
Vancouver, British Columbia, Canada
Sheraton Vancouver Wall Centre

UNESCO Memory of the World Programme, Knowledge Societies Division

This book of Proceedings includes most of the papers and posters presented at the International Conference "The Memory of the World in the Digital Age: Digitization and Preservation" held on 26–28 September 2012 in Vancouver, British Columbia, Canada, by the UNESCO Memory of the World Programme, Knowledge Societies Division, and The University of British Columbia in collaboration with the University of Toronto.

The proceedings have been compiled and formatted with minor editing; papers and posters appear as submitted. The authors are responsible for the choice and the presentation of the facts contained in this publication and for the opinions they express, which are not necessarily those of UNESCO and do not commit the Organization.

The designations employed and the presentation of material throughout this publication do not imply the expression of any opinion whatsoever on the part of UNESCO concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries.

The organizers of this UNESCO Memory of the World Programme Conference would like to sincerely thank everyone who contributed to the Conference in Vancouver and to these proceedings.

Published by UNESCO 2013, with the financial support of the Social Sciences and Humanities Research Council of Canada | Conseil de recherches en sciences humaines du Canada (SSHRC) and the International Research on Permanent Authentic Records in Electronic Systems (InterPARES) Project.

Social Sciences and Humanities Research Council of Canada | Conseil de recherches en sciences humaines du Canada | Canada

InterPARES Project
International Research on Permanent Authentic Records in Electronic Systems

Preface

This publication presents the proceedings of the international conference 'Memory of the World in the Digital Age: Digitization and Preservation' which was held in Vancouver, Canada, from 26 to 28 September 2012.

More than 500 experts and other interested persons from all regions of the world participated in this knowledge-sharing and policy-driving event to discuss and exchange opinions on how to protect the world's documentary heritage. Although this heritage is the record of knowledge, its physical carriers are extremely vulnerable and can easily disappear without a trace. Whether recorded on a clay tablet or an electronic tablet, our methods of sharing content and knowledge need to be protected.

It is impossible to exaggerate the importance of documentary heritage in our lives. It governs our actions whether these relate to creating the basis of mutual respect between different civilizations and communities or building knowledge societies. Documentary heritage provides the foundation of peace, our identity and knowledge.

UNESCO's interest in this subject matter is as fundamental as its constitution with its mandate to contribute to building peace through the spread of knowledge from improved access to printed and published materials. These core materials, our documentary heritage, have been preserved in archives, libraries and museums for generations.

But while measures needed to maintain access to print materials are globally understood, the newer challenges related to preserving digital information are not keeping pace with technological development. The need for dedicated hardware and software, associated with their rapid obsolescence, hamper our ability to keep invaluable content accessible. Unless timely migration to newer technologies, operating systems and software platforms is assured, we face the risk developing digital Alzheimer's.

UNESCO's expectation from this Conference was to obtain a better definition of our expected role, and our contribution to setting a global digital agenda. The UNESCO/UBC Vancouver Declaration sets out specific recommendations which we will be implementing and incorporating into our digital strategy. Likewise, we expect that our Member States, professional organizations and private sector bodies will also implement the recommendations addressed to them.

Only through collaborative strategic alliances can we overcome the major challenges threatening the preservation of digital information. We believe that the presentations featured in this publication provide the basis for a global commitment to preserving the memory of our world in this digital age.

Jānis Kārkliņš
Assistant Director-General
for Communication and Information

Contents

Preface 4

Opening Keynotes

Kenneth Thibodeau
Wrestling with Shape-Shifters: Perspectives on Preserving Memory in the Digital Age 15

Luciana Duranti
Trust and Conflicting Rights in the Digital Environment 24

Anne Thurston
Digitization and Preservation: Global Opportunities and Cultural Challenges 31

Intellectual Property Infrastructure Initiatives for Digital Heritage

Heather Christenson and John P. Wilkin
Intellectual Property Rights & the HathiTrust Collection 39

Elizabeth Townsend Gard
The Durationator® Copyright Experiment 46

Kate Hennessy
The Intangible and the Digital: Participatory Media Production and Local Cultural Property Rights Discourse 58

Preservation Infrastructures: Current Models and Potential Alternatives

Ilaria Pescini and Walter Volpi
An Example to Follow: An Infrastructure for Interoperability and Governance in the Tuscan Public System for Digital Preservation 70

Francis G. Mwangi
The Road to Providing Access to Kenya's Information Heritage: Digitization project in the Kenya National Archives and Documentation Service (KNADS) 83

Jeremy York
A Preservation Infrastructure Built to Last: Preservation, Community, and HathiTrust 92

Hrvoje Stančić, Arian Rajh and Ivor Milošević
"Archiving-as-a-Service": Influence of Cloud Computing on the Archival Theory and Practice 108

The CODATA Mission: Preserving Scientific Data for the Future

Elizabeth Griffin and the *CODATA* DARTG Team
Recovering the Forgettery of the World 127

Patrick C. Caldwell
Tide Gauge Data Rescue 134

Stephen Del Greco
Environmental Data Through Time: Extending The Climate Record 150

Tracey P. Lauriault and D. R. Fraser Taylor
The Map as a Fundamental Source in the Memory of the World 160

Preserving Tradition and Performing Arts in Digital Form

Ravi Katikala, Kurt Madsen and Gilberto Mincaye Nenquimo Enqueri
Life at the Edge of the Internet: Preserving the Digital Heritage of Indigenous Cultures 190

Lekoko Kenosi
Digital Madness, Archival Theory and the Endangered Sound Archives of Radio Botswana 206

Jørgen Langdalen
Editing Historical Music in the Age of Digitization 212

Lauren Sorensen and Tanisha Jones
Developing and Implementing a Digital Video Repository for Legacy Dance Documentation: Dance Heritage Coalition's Secure Media Network 217

Beyond Access: Digitization to Preserve Culture

Fernanda Maria Melo Alves, José António Moreiro González and José Manuel Matias
Safeguarding of the Portuguese Language Documentary Heritage: The Lusophone Digital Library 229

Benoit Ferland et Tristan Müller
Le réseau francophone numérique 236

John Van Oudenaren
The World Digital Library 246

Strategies for Building Digital Repositories

Bronwen Sprout and Sarah Romkey
A Persistent Digital Collections Strategy for UBC Library 257

Neil Grindley
Building the Business Case for Digital Preservation 269

Kevin Bradley
Requirements of a Remote Repository 278

Digital Forensics for the Preservation of Digital Heritage

Wayne W. Liu
Accountability for Archival Digital Curation in Preserving the Memory of the World 288

Christopher A. Lee and Kam Woods
Automated Redaction of Private and Personal Data in Collections: Toward Responsible Stewardship of Digital Heritage 298

Corinne Rogers and Jeremy Leighton John
Shared Perspectives, Common Challenges: A History of Digital Forensics & Ancestral Computing for Digital Heritage 314

Giving a Permanent Digital Voice to the Silenced

Terry Reilly
For the Children Taken: The Challenge to Truth Commissions in Building digital collections for research and long-term preservation 338

National Strategies as the Foundation of Togetherness

Andris Vilks and Uldis Zariņš
National Planning as the Key for Successful Implementation of Digitization Strategies 348

Ivan Chew & Haliza Jailan
Preserving the Crowdsourced Memories of a Nation: The Singapore Memory Project 354

Ernesto C. Bodê
Digital Preservation Policy of The Chamber of Deputies: Methodology for its development 366

Web 2.0 Products as Documentary Digital Heritage: Can We Access and Preserve Them?

Jamie Schleser
Unprotected Memory: User-Generated Content and the Unintentional Archive 378

Heather Ryckman
Context 2.0: User Attitudes to the Reliability of Archival Context on the Web 393

Lisa P. Nathan and Elizabeth Shaffer
Preserving Social Media: Opening a Multi-Disciplinary Dialogue 410

The Role of Culture in Digitization and Digital Preservation

Fiorella Foscarini, Gillian Oliver, Juan Ilerbaig and Kevin Krumrei
Preservation Cultures: Developing a Framework for a Culturally Sensitive Digital Preservation Agenda 419

Tukul Sepania Walla Kaiku and Vicky Puipui
Political, Cultural and Professional Challenges for Digitization and Preservation of Government Information in Papua New Guinea: An Overview 431

Xincai Wang and Yunxia Nie
Current Situation, Problems and Prospects of the Digital Preservation of Documentary Heritage in China 439

Open Archival Information System Reference Model: Answer or Inspiration?

Stefano S. Cavaglieri
Digital Archiving Systems Confronted with the OAIS Reference Model 451

Saeed Rezaei Sharifabadi, Mansour Tajdaran and Zohreh Rasouli
A Model for Managing Digital Pictures of the National Archives of Iran: Based on the Open Archival Information System Reference Model 457

Collaboration in Digital Preservation or Lack Thereof: What Works

Maria Guercio
Digital Preservation in Europe: Strategic Plans, Research Outputs and Future Implementation. The Weak Role of the Archival Institutions 467

Rolf Källman
Models for National Collaboration: Coordination of the Digital Cultural Heritage in Sweden 482

Victoria Reich
Building and Preserving Library Digital Collections Through Community Collaboration 489

Steve Knight
National Library of New Zealand, Digital Preservation and the Role of UNESCO 500

The Economics of Preserving Digital Information

David S. H. Rosenthal, Daniel C. Rosenthal, Ethan L. Miller, Ian F. Adams, Mark W. Storer and Erez Zadok
The Economics of Long-Term Digital Storage 513

Ulla Bøgvad Kejser, Anders Bo Nielsen and Alex Thirifays
Modelling the Costs of Preserving Digital Assets 529

L.M. Udaya Prasad Cabral
Economically Easy Method to Digitize Oversized Documents with Special Reference to Ola Leaf Manuscripts in Sri Lanka 540

Patricia Liebetrau
Preserving Our Heritage: An Independent Advantage 549

Is A New Legal Framework Required for Digital Preservation or Will Policy Do?

Tony Sheppard
Is a New Legal Framework Required for Digital Preservation or Will Policy Do? Building a Legal Framework to Facilitate Long-term Preservation of Digital Heritage: A Canadian Perspective 559

Alicia Barnard
Development of Policies and Requirements for Ingesting and Preserving Digital Records Into a Preservation System: Where to start? 570

Jason R. Baron and Simon J. Attfield
Where Light in Darkness Lies: Preservation, Access and Sensemaking Strategies for the Modern Digital Archive 580

Elaine Goh
Strengthening the Regulatory Framework in a Digital Environment: A Review of Archives Legislation 596

Digital Curation: Convergence of Challenges, Institutions and Knowledge

Sarah Higgins
Digital Curation: The Challenge Driving Convergence across Memory Institutions 607

Jackie R. Esposito
Digital Curation: Building an Environment for Success 624

Patricia Forget
Célébrations institutionnelles : Événement catalyseur propice à l'implantation d'un projet de conservation du patrimoine numérique permettant de réunir les acteurs d'intérêts divergents 636

Jeannette A. Bastian and Ross Harvey
The Convergence of Cultural Heritage: Practical Experiments and Lessons Learned 650

Digitization and Digital Preservation Experiences in a Developing Country Perspective

Elizabeth F. Watson
The Conservation and Preservation of Heritage in the Caribbean: What Challenges Does Digitization Pose? 661

Richard Marcoux, Laurent Richard and Mamadou Kani Konaté
Digital Preservation of Demographic Heritage: Population Censuses and Experiences in Mali and the Democratic Republic of the Congo 672

Brandon Oswald
Partnership in Paradise: The Importance of Collaboration for Handling Traditional Cultural Expression Material in the Pacific Islands 685

Ensuring That it Won't Happen Again

Victoria L. Lemieux
Financial Records and Their Discontents: Safeguarding the Records of our Financial Systems 700

Myron Groover
The White House E-Mail Destruction Scandal of 2007: A Case Study for Digital Heritage 713

Kenneth Thibodeau
The Perfect Archival Storm: The Transfer of Electronic Records from the G.W. Bush White House to the National Archives of the United States 724

Trusting Records

Lorraine Dong
The Ethical and Legal Issues of Historical Mental Health Records as Cultural Heritage 735

Marie Demoulin et Sébastien Soyez
L'authenticité, de l'original papier à la copie numérique : Les enjeux juridiques et archivistiques de la numérisation 745

Web Archiving as Part of Building the Documentary Heritage of Our Time

Liu Hua, Yang Menghui, Zhao Guojun and Feng Huiling
Chinese Web Archiving and Statistical Analysis on Chinese Web Archives 765

Gustavo Urbano Navarro
Implications of the Web Semantization on the Development of Digital Heritage 775

Matt Holden
Preserving the Web Archive for Future Generations: Practical Experiments with Emulation and Migration Technologies 783

Technology as the Mediator of Heritage and Its Relations with People

Ian S. King
The Turtle At The Bottom: Reflections on Access and Preservation for Information Artefacts 797

Erik Borglund
Challenges to Capture the Hybrid Heritage: When Activities Take Place in Both Digital and Non-Digital Environments 814

Limited Resources or Expertise: Case Studies in Addressing the Issue

Jean Bosco Ntungirimana
La problématique de la préservation de la mémoire collective au Burundi à l'ère des NTIC : Étude de cas menée à la Cour supreme 823

Farah Al-Sabah
Digitizing A Survivor's Identity: The Past, Present, and Future of the Kuwait National Museum Archives 838

Wayne W. Torborg, Theresa M. Vann and Columba Stewart
The Challenges of Manuscript Preservation in the Digital Age 851

Plenary 3 Keynotes

Dietrich Schüller
Challenges for the Preservation of Audiovisual Documents: A General Overview 863

International Perspectives and Cooperation

Claudia Nicolai, Rachele Oriente and Fernando Serván
One Year of Efforts for Digital Preservation at FAO 871

Peter Burnhill, Françoise Pelle, Pierre Godefroy, Fred Guy, Morag Macgregor and Adam Rusbridge
Archiving the World's E-Journals: The Keepers Registry as Global Monitor 880

The World Audiovisual Memory: Practical Challenges, Theoretical Solutions?

Jean Gagnon
Treasures That Sleep: Film Archives in the Digital Era 892

Caroline Frick
Seeing, Hearing, and Moving Heritage: Issues and Implications for the World's Audiovisual Memory in the Digital Age 896

Edoardo Ceccuti
The Digitization of Films and Photos of the Istituto Luce 904

Adam Jansen
Challenges and Triumphs: Preserving HD Video at the UBC School of Journalism 909

Mick Newnham, Trevor Carter, Greg Moss and Rod Butler
Digital Disaster Recovery for Audiovisual Collections: Testing the Theory 921

Metadata and Formats for Digitization and Digital Preservation

Joseph T. Tennis
Data, Documents, and Memory: A Taxonomy of Sources in Relation to Digital Preservation and Authenticity Metadata 933

Adam Rabinowitz, Maria Esteva and Jessica Trelogan
Ensuring a Future for the Past: Long-term Preservation Strategies for Digital Archaeological Data 941

Giovanni Michetti and Paola Manoni
It FITS the Cultural Heritage! Formats for Preservation: From Spatial Data to Cultural Resources 955

Lois Enns and Gurp Badesha
File Viewers: Examining On-the-Fly File Format Conversion 962

Walter Allasia, Fabrizio Falchi, Francesco Gallo and Carlo Meghini
Autonomic Preservation of "Access Copies" of Digital Contents 976

A Methodology Framework to Ensure Preservation

Anca Claudia Prodan
Bias and Balance in the Preservation of Digital Heritage 989

Giovanni Michetti
Archives Are Not Trees: Hierarchical Representations in Digital Environment 1002

Göran Samuelsson
The New Information Landscape: The Archivist and Architect – Drawing on a Common Map? 1011

Shadrack Katuu
Enterprise Content Management and Digital Curation Applications: Maturity Model Connections 1025

Christopher J. Prom
Facilitating the Aggregation of Dispersed Personal Archives: A Proposed Functional, Technical, and Business Model 1042

Digital Objects as Forensic Evidence

Carsten Rudolph and Nicolai Kuntze
Constructing and Evaluating Digital Evidence for Processes 1057

Aaron Alva, Scott David and Barbara Endicott-Popovsky
Forensic Barriers: Legal Implications of Storing and Processing Information in the Cloud 1064

Michael Losavio, Deborah Keeling and Michael Lemon
Models in Collaborative and Distributed Digital Investigation: In the World of Ubiquitous Computing and Communication Systems 1079

Fabio Marturana and Simone Tacconi
Cloud Computing Implications to Digital Forensics: A New Methodology Proposal 1093

Andrew F. Hay and Gilbert L. Peterson
Acquiring OS X File Handles Through Forensic Memory Analysis 1102

Institutional and Inter-Organizational Initiatives in Digitization

Anup Kumar Das
Digitization of Documentary Heritage Collections in Indic Language: Comparative Study of Five Major Digital Library Initiatives in India 1126

Ronald Walker
Digital Heritage Preservation - Economic Realities and Options 1139

S. K. Reilly
Positioning Libraries in the Digital Preservation Landscape 1146

Heidi Rosen, Torsten Johansson, Mikael Andersson and Henrik Johansson
Experiences from Digidaily: Inter-Agency Mass Digitization of Newspapers in Sweden 1153

Preserving Images: What Do We Need to Know?

Adama Aly Pam
Chemins de la mémoire : Les archives audiovisuelles au secours de l'identité d'une organisation internationale africaine 1163

Krystyna K. Matusiak and Tamara K. Johnston
Digitization as a Preservation Strategy: Saving and Sharing the American Geographical Society Library's Historic Nitrate Negative Images 1173

Jessica Bushey
Born Digital Images: Creation to Preservation 1189

Angelina Altobellis
Essential Skills for Digital Preservation: Addressing the Training Needs of Staff in Small Heritage Institutions 1198

Small and Large Scale Digitization: Towards a Shared Conceptual Model

Peter Botticelli, Patricia Montiel-Overall and Ann Clark
Building Sustainable Digital Cultural Heritage Collections: Towards Best Practices for Small-scale Digital Projects 1205

Marco de Niet, Titia van der Werf and Vincent Wintermans
Preserving Digital Heritage: The UNESCO Charter and Developments in the Netherlands 1219

Paul Conway
Validating Quality in Large-Scale Digitization: Findings on the Distribution of Imaging Error 1233

Lars Björk
Lost in Transit: The Informative Capacity of Digital Reproductions 1252

Preservation of Audiovisual Material

Mike Casey
The Media Preservation Initiative at Indiana University Bloomington 1266

George Blood
Video Compression…For Dummies? 1273

Pio Pellizzari, Álvaro Hegewich
The Ibero-American Preservation Platform of Sound and Audiovisual Heritage 1289

Trusting Data and Documents Online

Junbin Fang, Zoe Lin Jiang, Mengfei He, S.M. Yiu, Lucas C.K. Hui, K.P. Chow and Gang Zhou
Investigating and Analysing the Web-based Contents on Chinese Shanzhai Mobile Phones 1297

Junwei Huang, Yinjie Chen,Zhen Ling, Kyungseok Choo and Xinwen Fu
A Framework of Network Forensics and its Application of Locating Suspects in Wireless Crime Scene Investigation 1310

F.R. Van Staden and H.S. Venter
Implementing Digital Forensic Readiness for Cloud Computing Using Performance Monitoring Tools 1329

Yongjie Cai and Ping Ji
Security Monitoring for Wireless Network Forensics (SMoWF) 1340

Workshops

Peter Van Garderen, P. Jordan, T. Hooten, C. Mumma and E. McLellan
The Archivematica Project: Meeting Digital Continuity's Technical Challenges 1349

Hannes Kulovits, Christoph Becker and Andreas Rauber
Roles and Responsibilities in Digital Preservation Decision Making: Towards Effective Governance 1360

Posters and Presentations

Collence Takaingenhamo Chisita and Amos Bishi
Challenges and Opportunities of Digitizing and Preserving Cultural Heritage in Zimbabwe 1382

Donna McRostie
The long and winding road from aspiration to implementation – building an enterprise digitization capability at the University of Melbourne 1384

Asger Svane-Knudsen and Jiří Vnouček
Retrieving a part of Danish colonial history: From dust to digital copy 1386

Mitra Samiee and Saeed Rezaei Sharifabadi
A Paradigm for the preservation of national digital memory of Iran 1392

Chinyere Otuonye, Tamunoibuomi F. Okajagu, Samuel O. Etatuvie, Emmanuel Orgah, Gift Eyemienbai, Luke Oyovwevotu, Ewoma Borgu, and Janet Ukoha
Insights on the Digitization of Traditional Medicine Knowledge in Nigeria 1395

Nader Naghshineh and Saeed Nezareh
Crowd-sourced digital preservation: An Iranian model 1397

Chris Muller
Data at Risk: The Duty to Find, Rescue, Preserve 1399

Natalia Grincheva
Digital diplomacy: Providing access to cultural content, engaging audiences on a global scale 1401

Rusnah Johare
Preserving digital research data 1403

Claudia M. Wanderley
Multilingualism at the University of Campinas 1405

Anne Thurston
Open government and trustworthy records 1407

Jan Marontate, David Murphy, Megan Robertson, Nathan Clarkson and Maggie Chao
Canada – Aural memories: A case study of soundscape archives 1421

Na Cai, Leye Yao and Liu Liu
Creating Social Memories of Major Events in China: A Case study of the 5•12 Wenchuan Earthquake Digital Archive 1423

Addendum

Howard Besser
Archiving Large Amounts of Individually-Created Digital Content: Lessons from Archiving the Occupy Movement 1432

Nadja Wallaszkovits
Digitisation of Small Sound Collections: Problems and Solutions 1440

UNESCO/UBC Vancouver Declaration

The Memory of the World in the Digital Age: Digitization and Preservation 1452

Sponsors

File Viewers

Examining On-the-Fly File Format Conversion

Lois Enns[1] and Gurp Badesha[2]

[1]Records Manager, City of Surrey; [2]Functional Application Specialist, City of Surrey

Abstract

File viewers are utility applications that identify file formats and render source files in human-readable form using on-the-fly file format conversion and without triggering a native application. While many archives follow a file format conversion strategy for long-term digital preservation, other organizations may experience significant barriers to this preservation strategy in terms of resources, technology, risks, and drivers. Working within an InterPARES 3 general study, co-investigators at the City of Surrey tested six file viewer products to answer four research questions: 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. Based on test results, opportunities were identified for using file viewers as a component of a digital preservation strategy to reduce resource requirements, extend backwards compatibility, and improve electronic appraisal procedures.

Authors

Lois Evans has a Master of Information Studies from the University of Toronto, and has worked in local government for nine years as a district archivist and records manager, now with the City of Surrey. As a co-investigator on the UBC InterPARES 3 project, Lois developed an end-to-end procedure for appraising files on shared drives and migrating records to an electronic records management system, and published the *Shared Drive Migration Toolkit*. Lois has written on e-government and e-records for a number of publications.

Gurp Badesha has a Bachelor in Interactive Arts and Technology from Simon Fraser University, and has worked in local government at the City of Surrey for three years as a records coordinator and functional application specialist on the electronic records management system. Gurp participated in the shared drive migration project, and wrote the *Utility Applications Guide* which accompanies the *Shared Drive Migration Toolkit*.

1. Background

During the UBC InterPARES 3 case study on developing a production-oriented procedure for appraising and migrating files from shared drives to an electronic content management (ECM) system (Rogers *et al.*, 2010), the InterPARES co-investigators at the City of Surrey identified and adopted a number of utility applications to expedite our work. These utility applications included: a disk space manager, used to collect drive statistics, analyse file formats, create historical profiles, and facilitate metadata discovery; a file manager, used to apply unique identifiers and rename records; a duplication finder, used to identify and remove byte-by-byte duplicates; a format identifier, used to identify and resolve missing file extensions; and a empty folder identifier, used to count and remove empty folders. These activities are described in the *Shared Drive Migration Toolkit* (Enns and Badesha 2011). Although over 285 file formats were identified during the course of the project, only 47 file formats were confirmed as records suitable for migration, and only two of these file formats were found to be obsolete. These two file

Proceedings of The Memory of the World in the Digital Age: Digitization and Preservation. An international conference on permanent access to digital documentary heritage, 26-28 September 2012, Vancouver, British Columbia, Canada, edited by Luciana Duranti and Elizabeth Shaffer (UNESCO 2013).

formats (.ptn and .dwt) represented only 18 files out of 98,197 selected for migration. The remaining 45 file formats could be opened using available native applications.

The Surrey case study did not address digital preservation. Although many of the migrated records were scheduled for permanent retention, conversion to preservation formats was determined to be out of scope for the project, due to a number of barriers. Resource constraints included lack of disk storage space, staff capacity, staff time, and lack of documented standard operating procedures. Technical difficulties included a lack of capacity to manage bulk conversions and related metadata in either the source or target environments. Additionally, none of the conversion drivers identified in *ANSI/ARMA 16-2007 The Digital Records Conversion Process* (American National Standards Institute 2007)—such as retention requirements, operational factors, or regulatory or legal factors (p. 4)—appeared to fit the situation. Finally, the risk of "degradation or loss of the accuracy, completeness, authenticity, and integrity of the records" (p. 1) for format conversion appeared high, considering in the migration work already underway. For these reasons, both the records management and the information technology teams were reluctant to commit to a conversion strategy at this time.

As well, the records team were aware that the file formats in the Surrey environment appraised for migration were not necessarily subject to immediate technical obsolescence, since only two formats and 18 files were obsolete. Given that the vast majority of the files were not obsolete and did not appear to be under thread of obsolescence, the records team wondered whether the question of file conversion might be postponed indefinitely. Around the same time, the records team tested an ECM-integrated file viewer module that allowed users to open and annotate specialty drawing files (i.e., .dwg) without using the native application (i.e., AutoCAD). Although subsequent testing revealed that the module was not well-integrated to the ECM system (and it was not adopted), the idea that a file viewer might somehow extend the life of a file format was appealing.

As a secondary consideration, the records team found that during file appraisal activities, opening files to validate contents was a time-consuming activity. Only a few applications could be effectively managed on a computer task bar, and time was spent waiting for applications to open and files to load, and in flipping between native and utility applications. A file viewer supported multiple formats from a single point was worth pursuing.

In May 2011, the InterPARES 3 Team Canada members approved a general study on file viewers. 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 the next year, these questions were examined by the two Surrey con-investigators, with participation by two graduate research assistants, and input from members of Team Canada at bi-annual workshops. Study activities included: a literature review; correspondence with file viewer developers; selection of file viewer products for testing; development of basic product comprehension and creation of a test environment; identification of file formats, properties, characteristics, and files for testing; testing of products and collection of data; and examination of results.

2. 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 conferencing environment (Roseman and Greenberg 1995, p. 6). Other articles mention file viewers in the

context of software programming, along with 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). Evidently, 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, 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.*). Scattered through these 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.*) 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 a "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 conversion and/or pathway strategies for 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 [InSPECT], Creative Archiving at Michigan and Leeds Emulating the Old On the New [CAMiLEON], Consortium of Research Libraries Exemplars in Digital Archives [CEDARS], Preservation and Long-term Access through Networked Services [PLANETS]) and national archives (e.g., National Archives of Australia, National Archives and Records Administration [US], The National Archives [UK]) have published papers or web articles on significant properties, also called "significant characteristics" or "essential characteristics". Significant properties provide 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" which 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" (Lynch 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" (as quoted in Wilson 2007b, p. 4). However, there is an ensuing problem as what is considered "essential" varies 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. 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 referenced by their institutions on their websites.

3. Methods

Following the literature review, the co-investigators selected file viewer products for testing. Two categories of file-viewer software emerged: low-cost file viewers intended as stand-alone products; and more costly file viewers intended for integration with other software. This study focused on low-cost, stand-alone products costing less than $100 per license. Ease-of-use and the number of format categories covered by the product were two other important criteria. A number of Google searches were completed (e.g., "file viewers," "universal viewers," "best file viewers") and a preliminary list of products was identified.

Next, the products were qualified using Download.com, a site featuring software reviews, technology news and software downloads, and SouceForge.net, a site for open-source software development. Once the products were short-listed, each product website was reviewed to identify the best fit for the project, and the final product selection was made. Although open-source file viewers were identified, only one open-source file viewer supported two of the six format categories, and an attempt to download this product was unsuccessful due to programming requirements. In the end, the products selected for testing included: Accessory Software File Viewer ($23.00); FileStream Turbo Browser ($69.00); GetData Explorer View ($29.95); Irfan View ($10.00 donation); Quick View Plus ($49.00); and UV ViewSoft ($25.00). Once the products were selected, the co-investigators contacted the developers using email and web forums to ask questions about how file viewers work. In every case, the developers were advised that the co-investigators were seeking information for a research paper on file viewers. Most of the developers replied, and sufficient information was provided to create a general understanding of how file viewers work.

A test environment was set up to host the six file viewer products. The environment included two workstations: a Windows-platform workstation connected to Surrey's networked computing environment; and a Windows-platform personal laptop owned by one of the co-investigators and not connected to the network. All of the test files were maintained on the Surrey workstation, and all of the file viewers were downloaded to the personal laptop. The test files were transferred from the workstation to the personal laptop using a USB drive. Once the six file viewers were loaded to the laptop, the co-investigators spent some time orientating to the products, and eventually ran a complete set of test files to confirm their

understanding of the products and testing routine. The test run included seven file viewers (including one trial version later not adopted), 14 file formats, and nine files for each format, with three files selected from three time blocks (1994-1999; 2000-2005; and 2006-2011) to test whether file viewers are to any degree backwards compatible.

With the test environment and file viewers in place, the co-investigators looked for ways to measure how well the file viewers rendered files, referencing the significant properties listed on the InSPECT, Florida Digital Archives, and the Library of Congress websites for each format category. Here, the co-investigators took a somewhat different approach, separating significant properties into two somewhat arbitrary groups: **properties**, which could be determined without opening a file; and **characteristics**, which could only be determined by opening a file. For the purpose of this study, properties represent metadata that can be reviewed using a disk space manager, while characteristics represent metadata that cannot be viewed using the disk space manager as well as content. In a best-case scenario, the two groups would be separated into **metadata properties** and **content characteristics**, where properties include all metadata and characteristics reflect content alone.

For all format categories, three properties were consistently identified: Title, Creator, and Date Created. Additional properties were identified by file format category: Word Count (for text); Resolution, Bit Depth, Width, and Height (for images); and Length, Width, Height, Pixel Aspect Ratio, and Frame Rate (for moving images). These properties could be assessed using the Windows operating system and/or a file manager utility application, and the native application. Property data was collected and reviewed (see Table 3) but did not play a part in determining how well file viewers render files.

Characteristics that could be assessed using file viewers included: Header and Footer, Font Size and Colour, Images/Diagrams, Bullets and Numbering, Print, *Hyperlinks, Page Count, and Text Search* (for text); *Font Size and Colour,* Cells, *Formulas, Macros and Links, Frames/Page Breaks* (for data); Font Size and Colour, Sender, Receiver, Name, Date Sent, Date Received, Subject, Attachments, Body, Signature (for email); Division, Paragraph, Image, Link, Frame (for web); Font Size and Colour; Colour, Scalability, Sharpness, *Page Number* (for drawings); Colour, Completeness (for images); and Colour, Sound, and Back and Forward Navigation (for moving images). Mandatory characteristics (on which the later pass/fail assessments were made) are displayed in regular font, while *optional characteristics* are displayed in *italic* font. Some characteristics, such as slide presentation and animation (.ppt) or formulas, macros, and links (.xls) were not represented by any of the file viewers. The lack of conversion of these characteristics is also common to .pdf format conversion. These characteristics were treated as non-mandatory.

Of the 45 file formats migrated to the Surrey ECM system, only 12 file formats represented at least 500 files and up to 18 years worth of instances. These file formats became the focus of file viewer testing and included: .doc, .pdf, .ppt, .xls, .msg, .htm, .dwg, .vsd, .jpg, .tif, .mov, and .avi. While the selection of formats chosen by another organization might differ, the co-investigators felt that these formats were quite common, and represented formats they would need a file viewer to render if it was to be used in any appreciable way for production purposes. Once the formats were selected, significant care was taken to ensure that files chosen for testing presented the properties and characteristics of interest, and files were chosen from each of three time blocks (except for .avi, where only files from 2006-2011 were found). The testing was done twice, using two different sets of nine files for each of the 12 formats.

During preliminary testing, a discovery was made that four out of the six viewers could not render Microsoft files in the ".x" file formats (i.e., .docx, .pptx, .xlsx), designed to meet the Office Open XML standard. Additionally, the file viewers could open .htm files but 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 developers have not had time to bundle in an appropriate viewer. At any rate, these file formats were removed from the test sample.

In addition, two test files could not be opened in the native application and were considered corrupted. These files were removed and replaced.

Once the files were selected and placed on the workstation and the laptop, the six file viewers were tested. Each file was opened on the workstation using the native application, and then on the laptop using the file viewer. Using a file format instance chart (see Table 4, 5), each characteristic presented on the laptop was compared to the workstation, and given a pass or fail. In total, 72 file format instance charts were completed.

Using the file format instance charts, a determination was made as to whether or not the file viewer successfully rendered the file format for the time block. A pass meant that all mandatory characteristics were successfully rendered (see Tables 6, 7, 8). The formats were then grouped, and the file viewer was given a pass or fail for the format category (see Table 9).

4. Results

The results are presented with reference to the four research questions posed for the IP3 General Study.

4.1 How Do File Viewers Work?

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. Some file formats are rendered "as is" from the source file, while others are converted on-the-fly from the source format to a target format that can be rendered. In order to extend their file format rendering capabilities, file viewers often consist of a number of viewers bundled together. For example, one respondent noted their product used viewers from Internet Explorer (for text, html, and Microsoft Object Linking and Embedding or OLE files); Leadtools (for image files); and Delphi (for data files with open database compliancy or ODBC), while another product leveraged the Microsoft Internet Explorer engine (for html files); a doc-rtf converter (for text files); and Delphi (for data files). A third respondent referred to "third-party libraries," and a fourth noted the use of "outside-in" libraries which convert "foreign" formats to a generic format that leverages a standard viewer. As noted by one respondent, file viewers are "actually rendering a much smaller number of standard formats" than the 100 to 300 file formats commonly listed in their product information. The file viewer bundling approach was demonstrated during testing, when all six of the file viewers tested 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 only, while FileStream Turbo Browser is intended for wider use and extends to six format categories. During the product selection phase, the types of format categories targeted by the products were captured (see Table 1). Based on this product information, the co-investigators expected that the FileStream Turbo Browser and Quick View Plus viewers would perform the best during testing.

Table 1. File Viewer Rendering Capabilities by File Format Type (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

The more costly file viewers intended for integration with other software often offered additional features in concert with file rendering: format conversion; editing; annotation, redaction, and integration. These features were less common in the lower cost file viewers investigated in this study (see Table 2).

Table 2. File Viewer 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

4.2 What Software Is Available For Use?

As mentioned, available software falls into two categories: low-cost file viewers, intended as stand-alone products; and more costly file viewers, intended for integration with other software. The focus of this study was low-cost file viewers, and there are dozens products available, beyond the six products selected for this study.

4.3 How Accurately Do File Viewers Render Files?

As mentioned, a number of metadata properties were reviewed using a disk space manager and were not considered as part of the file viewer testing. These properties were collected in tables, reviewed, and set aside (see Table 3).

Table 3. File Format Properties (.doc).

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*

Next, file format characteristics of the files were compared using the native application rendering of the source file on the workstation, and the file viewer rendering of the file on the laptop. These results were recorded in 72 file viewer instants charts (i.e., six file viewers x 12 file formats). Nine files were tested for each format, with three files from each time period (i.e., 648 files). The characteristics were charted, with mandatory characteristics in regular font and *optional characteristics* in italics (see Table 4).

Table 4. File Viewer Instance Chart Showing Pass Results (.doc).

ACCESSORY SOFTWARE FILE VIEWER									
DOC		Header/ Footer	Font	Images/ Diagrams	Bullets	Hyperlink	Page Count	Text Search	Print
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

For each characteristic, the co-investors compared the native application rendering of a file with the file viewer rendering of the same file, and marked the file viewer characteristic with a pass or fail. Based on the mandatory characteristics, the file viewer passed (see Table 4) or failed (see Table 5). Although a file viewer was given a fail if just one mandatory characteristic failed, in most cases, the results of the test were fairly obvious, with a number of fails noted (see Table 5). If the characteristic was not present, it was marked as "N/A" (not applicable).

Table 5. File Viewer Instance Chart Showing Fail Results (.doc).

GETDATA EXPLORER VIEW								
DOC	Header/ Footer	Font	Images/ Diagrams	Bullets	Hyperlink	Page Count	Text Search	Print
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

The pass/fails for each file viewer and all 12 file formats were compiled into three charts, showing the performance of the file viewer on the three time blocks: newer files dated from 2006 to 2011; somewhat older files from 2000 to 2005; and older files from 1994 to 1999 (see Tables 6, 7, and 8).

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

File Viewer	DOC	PDF	PPT	XLS	MSG	HTM	DWG	VSD	JPG	TIF	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	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	TIF	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	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	TIF	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	FAIL	PASS	FAIL	PASS	FAIL	PASS	FAIL	FAIL	PASS	PASS	PASS	N/A

None of the file viewers successfully rendered all 12 file formats. Two file viewers were able to open 10 out of 12 formats: FileStream Turbo Browser and Quick View Plus. While Turbo Browser was unable to open .doc and .vsd files, Quick View was unable to open .mov or .avi files. Interestingly, all file viewers rendered the .jpg and .tif image files, for all three time blocks.

Using the File Viewer Capabilities by File Format charts, a final chart was created, indicating File Viewer Rendering Capabilities by File Format Type (see Table 9). A diagonal bar was used to indicate where the test results did not match expectations.

In terms of results, all file viewers successfully rendered at least one file format category. However, in the context of the Surrey testing and test environment, only Quick View Plus test results matched the expected results from product information. Overall, file viewers were more backward compatible than expected, and, in general, if a file viewer could render a file format, it could render older versions of the file format. (There were only two exceptions: 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.) Image file formats

Table 9. File Viewer Rendering Capabilities by File Format Type (based on testing)

Product	Text	Data	Email	Drawing	Images	Moving Images
Accessory Software File Viewer	No	Yes	No	No	Yes	No
FileStream Turbo Browser	No	Yes	Yes	No	Yes	Yes
GetData Explorer View	No	No	No	No	Yes	No
Irfan View	No	No	No	No	Yes	No
Quick View Plus	Yes	Yes	Yes	Yes	Yes	No
UV ViewSoft	No	Yes	No	No	Yes	Yes

(i.e., .jpg, .tif) were rendered by all six file viewers, while other formats were not rendered by a number of file viewers (i.e., .doc, .ppt, .msg, .dwg, .vsd, .mov).

In terms of results, all file viewers successfully rendered at least one file format category. However, in the context of the Surrey testing and test environment, only Quick View Plus test results matched the expected results from product information. Overall, file viewers were more backward compatible than expected, and, in general, if a file viewer could render a file format, it could render older versions of the file format. (There were only two exceptions: 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.) Image file formats (i.e., .jpg, .tif) were rendered by all six file viewers, while other formats were not rendered by a number of file viewers (i.e., .doc, .ppt, .msg, .dwg, .vsd, .mov).

File viewers also demonstrated the conventional data/content loss limitations commonly noted in .pdf conversions, namely that: formulas were not displayed (.xls); slide presentation and animation was missing (.ppt); and hyperlinks did not work (.doc). This makes sense, as a number of the file viewers used the Adobe Acrobat viewer for on-the-fly conversion as well as .pdf rendering.

In fact, measuring how well the six file viewers rendered files made the co-investigators more aware of .pdf format limitations. There are many benefits to using .pdf as a preservation format: an open standard; a strong working group; a new version in development (i.e., PDF Universal Access, or PDF/UA); a fixed form that is portable, reliable, and interoperable; and billions of instances in existence. However, there are challenges in using the .pdf format for file format conversion. These are challenges are outlined (and debated) in a blog thread entitled, “After Flash, PDF Must Die” (Huber, 2012) and include: a non-reversible transformation requiring the preservation of native files; content/data loss (e.g., formulas, presentation, animation, hyperlinks); and tagging requirements (i.e., to optimize retrieval or re-use across devices). An interesting argument is made that the .pdf format may be “the software version of microfiche,) and that in the future, libraries will need to implement .pdf readers to provide access to the billions of files being created today. Time will tell, although it is interesting to note that the .tif format was seen as a de facto preservation format through the 1990s and early 2000s and now is regularly passed over in favour of the .pdf format. This discussion is continued later on, in the context of non-reversible transformation.

4.4 What Role Might File Viewers Play in Digital Preservation?

File viewers do allow rendering of file formats on-the-fly, with results similar to digital conversion and without the some of the resource requirements, technical difficulties, or migration risks. For the file formats selected in the general study, the file viewers proved to be backward compatible, and able to render files over an 18-year period, without accessing the native applications. File viewers are useful appraisal tools, as files can be rendered without opening native applications which can be difficult to effectively manage during appraisal activities. In some environments, file viewers enable access where native applications that are not resident in the appraisal environment, and also alleviate software licensing costs. For these reasons, file viewers may be considered by some organization to be a viable tool, or even a component of a digital preservation strategy.

File viewers do not overcome problems associated with content/data loss but do underline the somewhat overlooked problem of non-reversible transformation. Although some researchers believe that digital objects "do not need to remain in a state that is unchanged" (Wilson, 2007b, p. 4), researchers on the CAMiLEON project participants 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). The standard for preservation conversion was reversible transformation, as "the only way of ensuring a migration step has been completed without error is by the proof of reversible migration" (p. 519).

The problem with both preservation migration and file viewer conversion is that content is often lost through the representation of the native byte stream in the new format. Through this examination of file viewers, the most important consideration was how to assess the file viewers in terms of properties and characteristics. Depending on the expectations for properties and characteristics, test results would change so that more file viewers might "fail" or "pass." Although properties, in the sense of file property metadata, are clearly conveyed through standards, data dictionaries, and many other forums, characteristics are more difficult to assess, and further work is likely needed. Based on this study, there are at least three categories of characteristics that are important for assessing file format conversion: 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 were noted in the *InSPECT Significant Properties Report* (Wilson, 2007b), with reference to content, context, appearance, structure, and behaviour.

6. Conclusion

In closing, the co-investigators recognize the migration-conversion approach as the primary digital preservation strategy in place in archives 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, the migration-conversion strategy is not perfect, as characteristics are often lost during file transformations. With many institutions maintaining the native files in addition to a preservation copy, opportunities exist to pursue complementary strategies. For these reasons, the co-

investigators suggest that file viewers provide an opportunity to leverage native files in a digital preservation strategy. Here, the co-investigators note an extensive body of work on file formats in progress beyond the field of archives and records management, and the need to collaborate with these other fields of study, including software development.

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 the literature review for the *General Study on File Viewers,* the contributions of Dr. Luciana Duranti, InterPARES Project Director, and the input of Team Canada members.

References

Accessory Software File Viewer 9. Accessed February 23, 2012. http://www.accessoryware.com/FileView.htm.

Amirani, M. C., M. Toorani, and A. A. B. Shirazi. "A New Approach to Content-Based File Type Detection." In *Proceedings of the 13th IEEE Symposium on Computers and Communications (ISCC 2008), July 2008*, 700-705.

Anderson, P., and T. Teitelbaum. "Software Inspection Using Codesurfer." In *Proceedings of the First Workshop on Inspection in Software Engineering (WISE 2001), Paris. July 2001*, 1-9.

ARMA International. "The Digital Records Conversion Process: Program Planning, Requirements, Procedures." USA: ARMA International, 2007.

CNET Download.com. Accessed February 23, 2012. http://www.download.com.

Daeja ViewONE. Accessed February 23, 2012. http://www.daeja.com/products/viewone-pro-overview.asp.

Dhanalakshmi, R., and C. Chellappan. "File Format Identification and Information Extraction." In *Proceedings of the2009 World Congress on Nature and Biologically Inspired Computing (NaBIC 2009), Coimbatore, India, 9-11 December 2009*, edited by Ajith Abraham, Andre Carvalho, Francisco Herrera and Vijayalakshmi Pai, 1497-1501. 2009. http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=5393688.

Enns, L., and G. Badesha. *Shared Drive Migration Toolkit*. Canada: City of Surrey, 2011.

FileStream Turbo Browser. Accessed February 23, 2012. http://www.filestream.com/turbobrowser/.

The Florida Centre for Library Automation. "Florida Digital Archives." 2009. Accessed February 23, 2012. http://fclaweb.fcla.edu/FDA.

GetData Explorer View. Accessed February 23, 2012. http://www.explorerview.com/.

Grace, S., G. Knight, and L. Montague. *Investigating the Significant Properties of Electronic Content Over Time: Final Report*. United Kingdom: Centre for e-Research, King's College London, 2009. http://ie-repository.jisc.ac.uk/526/1/InSPECT_Final_Report_v1.pdf.

Huber, Serge. "After Flash, Why PDF Must Die!" *Social Business Expert Blog*. April 12, 2012. Accessed May 25, 2012. http://www.aiim.org/community/blogs/expert/after-flash-why-pdf-must-die-!.

Knight, G. *InSPECT Investigating Significant Properties of Electronic Content*. United Kingdom: Centre for e-Research, King's College London, 2007. http://www.significantproperties.org.uk/.

Knight, G. *Investigating the Significant Properties of Electronic Content Over Time: Framework Report*. United Kingdom: Centre for e-Research, King's College London, 2009. http://www.significantproperties.org.uk/inspect-framework.html.

Library of Congress. "Sustainability of Digital Formats: Planning for Library of Congress Collections." 2010. Accessed February 23, 2012. http://www.digitalpreservation.gov/formats/.

Mancoridis, S., R. C. Holt, and D. A. Penny. "A 'Curriculum-Cycle' Environment For Teaching Programming." In *Proceedings of the 24th SIGCSE Technical Symposium on Computer Science Education* (*SIGCSE '93). New York, 1993*.

McDaniel, M., and M. H. Heydari. "Content Based File Type Detection Algorithms." In *Proceedings of the 36th Hawaii International Conference on System Sciences (HICSS'03)*. 2002. http://dl.acm.org/citation.cfm?id=821828.

McHenry, K., R. Kooper, and P. Bajcsy. "Towards a Universal, Quantifiable, and Scalable File Format Converter." In *Fifth IEEE International Conference on e-Science. Oxford, December 9-11, 2009*. http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=5380873.

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

Quick View Plus. Accessed February 23, 2012. http://www.avantstar.com/metro/home/Products/QuickViewPlusStandardEdition.

Rogers, C., S. Malmas, and L. Enns. "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." Vancouver: InterPARES 3 Project, 2011.

Roseman, M., and S. Greenberg. "Building Real-Time Groupware with GroupKit, A Groupware Toolkit." *ACM Transactions on Computer-Human Interaction* 3, no. 1 (1995): 1-30.

Skiljan, I. "Irfan View." 2005. Accessed February 23, 2012. http://www.irfanview.ca/.

Sourceforge. Accessed February 23, 2012. http://sourceforge.net/.

Stellant Outside In. Accessed February 23, 2012. http://www.oracle.com/us/technologies/embedded/025613.htm.

Universal Viewer. Accessed February 23, 2012. http://www.uvviewsoft.com/.

Wilson, A. (2007a). "The Significant Properties of Digital Objects." JISC Significant Properties Workshop, British Library, April 2008. http://www.jisc.ac.uk/whatwedo/programmes/preservation/2008sigprops.

Wilson, A. (2007b). "InSPECT: Significant Properties Report." *Arts and Humanities Data Service*, 10 April 2007. http://www.significantproperties.org.uk/wp22_significant_properties.pdf.

Gold Sponsors

McGill

InterPARES Project
International Research on Permanent Authentic Records in Electronic Systems

Silver Sponsors

Ministry of Education, Culture and Science

OCUL Ontario Council of University Libraries

UNIVERSITY OF ALBERTA LIBRARIES

uOttawa Library

Supporters

UNESCO CANADIAN COMMISSION FOR UNESCO
COMMISSION CANADIENNE POUR L'UNESCO
www.unesco.ca

Internet Society

Manitoba

OCLC

Québec

uOttawa École des sciences de l'information
School of Information Studies