FINAL PROGRAM AND PROCEEDINGS

ARCHIVING2018

April 17-20, 2018 • Washington, DC

General Chair: Don Williams, Image Science Associates

Sponsored by the Society for Imaging Science and Technology

www.imaging.org/archiving

IS&T
imaging.org

The papers in this volume represent the program of Archiving 2018,
held April 17–20, 2018, at the National Archives and Records Administration (NARA) in Washington, DC.

Society for Imaging Science and Technology
7003 Kilworth Lane • Springfield, VA 22151 USA
703/642-9090; 703/642-9094 fax
info@imaging.org; www.imaging.org

ISBN Abstract Book: 978-0-89208-332-9
ISBN USB Stick: 978-0-89208-333-6
ISSN Print: 2161-8798
ISSN Online: 2168-3204

Manuscripts are reproduced from PDFs as submitted and approved by authors; no editorial changes have been made.

Printed in the United States.

Conference Sponsor

DIGITAL TRANSITIONS
DIVISION OF CULTURAL HERITAGE

Conference Exhibitors

Archeio

ColorBurst®

CROWLEY

HASSELBLAD
CREATE TO INSPIRE

PICTURAE

startext
innovation in software

Conference Partner

NATIONAL ARCHIVES

WELCOME TO ARCHIVING 2018

It's Springtime in Washington. Speaking for the program's organizing committee, we're happy to have you here for the 15th annual IS&T Archiving conference, held for the third time at the US National Archives and Records Administration (NARA). While I'm sure you'll find plenty of enjoyable experiences in the nation's capital, let me briefly share with you a summary of the conference program itself.

We've continued with some of the past popular short courses and also added several suggested new ones. We're trying to stay relevant based on past years' feedback. The instructors in the courses are very good at what they do and can offer both technical and workflow suggestions on their subject matter.

The four Keynote speakers will speak on diverse archiving subjects. I've listed them below to insure you don't miss them.

- Montreux Jazz Digital Project: From a Patrimony to an Innovation Platform, Alain Dufaux, EPFL Metamedia Center (Switzerland)
- 30 Years of 3D – Next Steps for Archiving a Disappearing World, Alonzo Addison (US)
- Enhancing Access to Collections, Partnering with the Public and Enriching the Museum and Archives Fields: The Robert F. Smith Fund at the National Museum of African American History and Culture, Doretha Williams, National Museum of African American History and Culture (US)
- Sound Preservation: Not Fast-Enough-Forward, Sam Brylawski, University of California, Santa Barbara (US)

The technical sessions are a nice mix of image digitization, data management and access, as well as archiving subjects by contributors from a range of countries. Don't miss the interactive sessions though. These tend to be presentations that are not done justice with an oral presentation and, to be appreciated, as the name implies are better suited to one-on-one personal discussions with the authors.

Be sure to visit and explore the offering of the exhibitors. They're likely to have new resources, tools, and products that can make your work life easier. At the rate technology is changing, advances can occur quickly.

Finally, definitely try to participate in one of the tour offerings. They're pretty neat. Even if some of the tours are filled don't be shy in visiting some of the many other free museum and gallery offerings in Washington. My favorites are the National Portrait Gallery and the National Building Museum, but that's just me. Explore and enjoy. Washington's Metro transportation services are easy to understand and travel. So, have fun and don't forget to make new professional contacts and renew old ones.

—Don Williams
General Chair, Archiving 2018

CONFERENCE COMMITTEE

General Chair
Don Williams, Image Science Associates (US)

Program Chair
Lukas Rosenthaler, University of Basel (Switzerland)

Short Course Chair
Jeanine Nault, Library of Congress (US)

Technical Program Committee
Michael J.Bennett, University of Connecticut (US)
Erik Landsberg, Cultural Heritage Digitization Consulting (US)
Martina Hoffmann, National Library of the Netherlands (the Netherlands)
Raivo Ruusalepp, National Library of Estonia (Estonia)
Christoph Voges, Hochschule für angewandte Wissenschaft und Kunst (HAWK), and consultant (Germany)

Steering Committee
Ulla Bøgvad Kejser,
Det Kongelige Bibliotek/The Royal Library (Denmark)
Peter Burns, Burns Digital Imaging (US)
Suzanne E. Grinnan, IS&T (US)
Lukas Rosenthaler, University of Basel (Switzerland)
David Walls, US Government Publishing Office (US)
Don Williams, Image Science Associates (US)

Paper Reviewers
Stephen Abrams, California Digital Library (US)
Michael J. Bennett, University of Connecticut (US)
Roy Berns, Rochester Institute of Technology (US)
Robert Buckley, National Archives of the UAE (UAE)
Peter Burns, Burns Digital Imaging (US)
Noah Durham, National Archives and Records Administration (US)
Roger Easton, Rochester Institute of Technology (US)
Peter Fornaro, University of Basel (Switzerland)
Rebecca Frank, University of Michigan School of Information (US)
Steffen Hankiewicz, intranda GmbH (Germany)
Lei He, Library of Congress (US)
Julia Hickey, Defense Media Activity (US)
Martina Hoffmann, National Library of the Netherlands (the Netherlands)
Michael Horsley, National Archives and Records Administration (US)
Ulla Bøgvad Kejser, Det Kongelige Bibliotek/The Royal Library (Denmark)
Erik Landsberg, Cultural Heritage Digitization Consulting (US)
Anne Mason, National Archives and Records Administration (US)
Phil Michel, Library of Congress (US)
Jeanine Nault, Library of Congress (US)
Jonas Palm, Swedish National Archives (Sweden)
Kristen Ratanatharathorn,The Andrew W. Mellon Foundation (US)
John Rees, National Library of Medicine (US)
Thomas Rieger, Library of Congress (US)
Lukas Rosenthaler, University of Basel (Switzerland)
Raivo Ruusalepp, National Library of Estonia (Estonia)
Richard Solomon, University of Pennsylvania (US)
Dave Thompson, Wellcome Library (UK)
Marie Vans, HP Inc. (US)
Christoph Voges, Hochschule für angewandte Wissenschaft und Kunst (HAWK), and consultant (Germany)
David Walls, US Government Publishing Office (US)

Archiving 2018 Cooperating Societies

- AIC American Institute for Conservation Foundation of the American Institute for Conservation
- ALCTS Association for Library Collections & Technical Services
- CNI Coalition for Networked Information
- Digital Library Federation at CLIR
- DPC Digital Preservation Coalition
- IOP Printing and Graphic Sciences Group
- ISCC Inter-Society Color Council
- MCN Museum Computer Network
- RPS The Royal Photographic Society

IS&T BOARD OF DIRECTORS
JULY 2017 – JUNE 2018

President
Steven Simkse, Colorado State University

Executive Vice President
Scott Silence, Corning Inc.

Conference Vice President
Francisco Hideki Imai, Apple Inc.

Publications Vice President
Susan Farnand, Rochester Institute of Technology

Secretary
Dietmar Wueller, Image Engineering GmbH & Co. KG

Treasurer
Eric Hanson, retired, HP Laboratories

Vice Presidents
Sergio Goma, Qualcomm Technologies, Inc.
Robin Jenkin, NVIDIA
Teruaki Mitsuya, Ricoh Company, Ltd.
James Stasiak, HP Inc.
Radka Tezaur, Intel Corporation
Werner Zapka, XaarJet Ltd. Filial

Immediate Past President
Geoff Woolfe, retired, Canon Information Systems Research Australia Pty. Ltd.

Chapter Directors
Korea: Choon-Woo Kim, Inha University
Rochester: David Odgers, Odgers Imaging
Tokyo: Masahiko Fujii, Fuji Xerox Co., Ltd.

IS&T Executive Director
Suzanne E. Grinnan

TECHNICAL PAPERS PROGRAM:
CONFERENCE SCHEDULE AND TABLE OF CONTENTS*

TUESDAY APRIL 17, 2018

Archiving 2018 Short Course Program

8:00 – 10:00 (2 hours)
ArchSC01: Spectral Imaging—Spectral Data and Technical Aspects
Instructors: Fenella France and Meghan Wilson, Library of Congress
ArchSC02: Scanner & Camera Imaging Performance: Ten Commandments
Instructors: Don Williams, Image Science Associates, and Peter Burns, Burns Digital Imaging

8:00 – 12:00 (4 hours)
ArchSC03: An Introduction to Digital Archiving
Instructor: John Sarnowski, ResCarta Foundation
ArchSC04: Preservation Strategies for Computational Photography based Imaging: Reflectance Transformation Imaging (RTI) and 3D Photogrammetry
Instructors: Carla Schroer and Mark Mudge, Cultural Heritage Imaging

10:15 – 12:15 (2 hours)
ArchSC05: Spectral Image Processing
Instructors: Fenella France and Meghan Wilson, Library of Congress
ArchSC06: Quality Assurance Workflows for Digitization Projects
Instructor: Martina Hoffmann, National Library of the Netherlands (KB)

13:30 – 15:30 (2 hours)
ArchSC07: Color Measurement for Archiving
Instructor: David R. Wyble, Avian Rochester, LLC
ArchSC08: Metadata and Workflows for DAMS: Building Blocks to Access
Instructors: Stephanie Christensen and Isabel Meyer, The Smithsonian Institution

13:30 – 17:45 (4 hours)
ArchSC09: Management of Multispectral and Advanced Image Data
Instructor: Michael B. Toth, R.B. Toth Associates
ArchSC10: Digital Audiovisual File Formats: Identification, Validation, Specification Verification
Instructors: Ashley Blewer, consultant, and Julia Kim, Library of Congress

15:45 – 17:45 (2 hours)
ArchSC11: Introduction to Color Management for Cultural Image Capture
Instructors: Don Williams, Image Science Associates, and Peter Burns, Burns Digital Imaging
ArchSC12: Unlocking the Power of (Linked) Metadata
Instructor: Martijn van der Kaaij, Heron Information Management LLP

Archiving 2018 Welcome Gathering

17:30– 19:30
Iron Horse Tap Room, 507 7th St. NW
Join colleagues following the short course program. Iron Horse Tap Room is a short walk from NARA.

WEDNESDAY APRIL 18, 2018

Welcome Remarks and Opening Keynote

Session Chairs: Don Williams, Image Science Associates (US), and Lukas Rosenthaler, University of Basel (Switzerland)
9:00 – 10:00

Montreux Jazz Digital Project: From a Patrimony to an Innovation Platform, *Alain Dufaux, EPFL Metamedia Center (Switzerland)*
Since 1967, audiovisual recordings of the Montreux Jazz Festival bring together the greatest musicians of the 20th century. The collection was inscribed on the 2013 UNESCO memory of the world register. Over 5,000 hours of 'live' concerts were recorded in state-of-the-art broadcast quality for both video and audio, of which a large part exists as multi-tracks.

The collection was digitized in a collaboration between EPFL and the Claude Nobs Foundation. The Montreux Jazz Digital Project aims to preserve and transform this heritage into a unique archive of "raw material" for researchers to innovate in the field of music technology, signal processing, acoustics, multimedia, design and even architecture. Adding value to the collection, a substantial metadata enrichment program will be devised for schools, musicians, and musicologists. In the recently built Montreux Jazz Café at EPFL, innovative user-interaction tools are placed at the archive's disposal to transform it into a living collection.

New Digitization Methods

Session Chair: Peter Fornaro, University of Basel (Switzerland)
10:00 – 12:15

10:00 **New Techniques for the Digitization of Art Historical Photographic Archives—the Case of the Cini Foundation in Venice,**
Benoit Seguin, Lisandra Costiner, Isabella Di Lenardo, and Frédéric Kaplan, École Polytechnique Fédérale de Lausanne (Switzerland) . . . **1**
Numerous libraries and museums hold large art historical photographic collections, numbering millions of images. Because of their non-standard format, these collections pose special challenges for digitization. This paper address these difficulties by proposing new techniques developed for the digitization of the photographic archive of the Cini Foundation. This included the creation of a custom-built circular, rotating scanner. The resulting digital images were then automatically indexed, while artificial intelligence techniques were employed in information extraction. Combined, these tools vastly sped processes which were traditionally undertaken manually, paving the way for new ways of exploring the collections.

10:25 – 11:00 Morning Coffee Break / Exhibit Open

* Page numbers indicate the page on which the paper is found in the full version of the conference proceedings, found on the accompnaying USB stick.

11:00 **3D Scanning Solution for Textured Object using Photometric Stereo with Multiple Known Light Sources,** *Arnold Cheveau, i2S (France)* **6**

Photometric Stereo is an efficient image-based 3D reconstruction technique that has been used to reproduce very high-quality reconstructions. However, it faces a couple of limitations: first, one needs to capture images of the 3D scene with different illumination directions. It implies that the 3D scene remains motionless during illumination changes, which prevents the reconstruction of deforming objects. Second, the captured images must be obtained from a single point of view. This leads to depth-map based 2.5D reconstructions, instead of full 3D surfaces.

But compared to other 3D imaging methods such as geometry modeling and 3D-scanning, this solution is a valuable tool when examining embossed surfaces where grain texture, carving, deteriorations can be identified.

In this paper, we give an outline of Photometric Stereo and provide a case study of our 3D scanner.

11:25 **Digitizing and Managing 35mm Mounted Slides: The Flip Side,** *Benjamin Sullivan and Walter Larrimore, Smithsonian Institution, National Museum of African American History and Culture (US)* **10**

Cultural heritage organizations of all types and sizes commonly maintain and preserve collections of 35mm mounted slides, oftentimes numbering in the hundreds, to thousands, to hundreds of thousands. Digitization of these objects presents multiple challenges. The mutual dualities of frontside/backside, combined with simultaneous reflective/transmissive content capture requires unusual imaging equipment and techniques to create efficient rapid capture workflows to meet current cultural heritage archival documentation requirements at scales such as these. Further, the interpretation, creation, and archiving of metadata from such captures present concomitant challenges, which may often be best met by integration into the imaging and processing workflow at the time of capture. Our research and development project created a suite of workflows and protocols for efficient and safe handling of slides as museum objects, complete data capture with current digital imaging studio equipment, and efficient post-processing of the digital image files.

11:50 **Digitizing Braille Music: A Case Study,** *Donna Koh and Katherine Rodda, Library of Congress (US)* **21**

The Music Section at the National Library Service for the Blind and Physically Handicapped at the Library of Congress has been digitizing its tactile braille music collection in order to preserve it, make it electronically available, and to reduce physical space needed for storage. Poor scanning and editing can result in scores that are confusing or even unusable, especially for the blind musicians who rely on our materials. Over the past few years, we have used different scanners and software with varying degrees of accuracy and speed. In this paper, we will explain our digitization process, the types of software and techniques we use, and discuss the challenges we face in capturing and proofreading archival quality e-braille files.

One Interactive Preview and 2-Minute Exhibitor Previews

Session Chair: Peter Fornaro, University of Basel (Switzerland)

12:15 – 12:30

Digal vs. Analogous Long Term Preservation. Microfilm Still Alive . . . ? (Interactive Preview), *Michael Luetgen, Zeutschel GmbH (Germany)* . . . **26**

Please note that this author will only be available to discuss his Interactive (Poster) Paper during the Wednesday afternoon coffee break.

The microfilm as a medium for long term preservation is still alive. Especially in the archives the microfilm is part of their strategies. But also libraries are using microfilm until today - also it's not a user friendly media type and access to information is very limited and uncomfortable.

The goal of this paper is to give an overview about the current status of analogous technology and analogous Long Term Preservation (examples, standards and tendencies), current status of digital Long Term Preservation, analogous equipment, risk management, cost comparison digital vs. analogous, resume and practical hints.

This paper will use experiences mostly from German examples but also international experiences from point of view of a vendor.

Archiving 2018 exhibitors Archeio, ColorBurst, Crowley, Hasselblad, Picturae, and startext share information about their products and services in 2-minute previews.

12:30 – 14:00 Lunch Break

Afternoon Keynote

Session Chair: Don Williams, Image Science Associates (US)

14:00 – 14:50

14:00 **30 Years of 3D—Next Steps for Archiving a Disappearing World,** *Alonzo Addison (US)*

It has been almost 3 decades since the advent of 3D digital documentation in the heritage domain. From photogrammetry to laser scanning and more, today's high-tech sensors allow us to rapidly record everything from great monuments to museum masterpieces, and precious manuscripts to intangible traditions. Across the globe, institutions, researchers, and even the public are adding terabytes of 3D data to archives and collections by the day. Yet capturing reality in digital form is only one step in a complex process. Sadly the majority of this data will not outlive the heritage it seeks to help conserve. In the rush to digitally preserve the past in 3D, we lack a coordinated plan and strategy. With examples from the advent of terrestrial lidar, to international initiatives in heritage policy, we will explore the pitfalls and potential for archiving a disappearing world.

Guidelines, Standards

Session Chair: Martina Hoffmann, National Library of the Netherlands (the Netherlands)

14:50 – 16:05

14:50 **Digitization with Use of Principles from the World of Industry,** *Marc Holtman and Nelleke van Zeeland, City Archives Amsterdam (the Netherlands)* **29**

"All archival research should be possible 24/7."

In 2005, the Amsterdam City Archives set its digital services department the ambitious target of making it possible to access its entire

archive at all times. This meant that we needed to start digitizing our archives and collections on a large scale.

An important part of our approach is customer-driven: through a scanning-on-demand service, the customer decides which archival documents are digitized. Next to scanning-on-demand, we work on a project basis in which we digitize entire archives or collections, of larger amounts of uniform material.

Ten years on, we have met 40,000 client orders and produced 20 million scans, all online for everyone to use. Meanwhile, the customer demand for digitized items continues to increase. At the same time, large digitization projects come our way more often. This requires constant adjustment of our digitization approach and work processes.

In 2018, we meet a new challenge: producing 20,000 scans a day. On this scale, an industrial approach as described in this article is not an option anymore, but a necessity.

15:15 **Developing Guidelines for Digitization of US Federal Government Records,** *Michael Horsley and Kevin L. DeVorsey, National Archives and Records Administration (US)* **33**

The United States National Archives and Records Administration (NARA) has issued a new Strategic Plan FY 2018-FY 2022 that reflects NARA's ongoing commitment to support the transition to electronic record keeping, increased access and a fully digital government. The plan provides federal agencies with a deadline: "By December 31, 2022, NARA will, to the fullest extent possible, no longer accept transfers of permanent or temporary records in analog formats and will accept records only in electronic format and with appropriate metadata."

It is likely that Agencies will digitize large volumes of paper records to meet this goal. This paper presents NARA's Records Management Policy and Standard's Team's draft guidance on digitizing records. The paper discusses the evolution of NARA's 2004 digital imaging for access guidelines into the current 2016 FADGI guidelines, and discusses some of the unique issues applying these guidelines in a records management context to comply with NARA's Strategic Plan.

15:40 – 16:15
Coffee Break, Interactive Paper Discussions, Exhibit Open

16:15 **Into the Deep: Adapting ISO Methods for Measuring Depth-of-Field,** *Don Williams, Image Science Associates, and Peter D. Burns, Burns Digital Imaging (US)* **37**

Monitoring of imaging performance is well-established and the subject of both imaging standards and guidelines for cultural heritage institutions. To date emphasis has been on the imaging of flat objects. As more three-dimensional content is being captured though, performance metrics for this class of materials need to be introduced and considered. Chief among these is depth-of-field (DOF), the distance of acceptable focus along the optical axis in front of the lens. We propose adapting image-resolution tools for arriving at a practical method for measuring depth-of-field. We discuss requirements for test-chart objects and analysis software.

Multispectral & 3D I

Session Chair: Fenella France, Library of Congress (US)

16:40 – 17:30

16:40 **Spectral Implications for Camera Characterization Target,** *David Wyble, Avian Rochester, LLC (US)* **42**

Accurate camera calibration is a critical step in the capture, processing, and archiving of object properties. To be most useful to the library/museum/archiving community, the patch colors in a camera color characterization target should facilitate accurate data capture from commercial RGB cameras. Target patches can be defined colorimetrically (*i.e.:* CIELAB) or spectrally (*i.e.:* reflectance). For some limited situations, colorimetric data is sufficient, but knowing and using the spectral reflectance of the patches affords increased flexibility and accuracy. In this work, the spectral reflectance of the patches are considered in light of the spectral detection properties of cameras. A spectral model will be developed to predict how well two commercial cameras perform when profiled against an available camera target.

17:05 **Practical UV-VIS-NIR Multispectral Imaging,** *Roy S. Berns, Rochester Institute of Technology (US)* **47**

A seven-channel multi-spectral camera has been developed using commercial products and both commercial and custom software. The camera components are manufactured by Finger Lakes Instrumentation, Rodenstock, and the Andover Corporation. The colored glass filters were optimized for image quality, colorimetric accuracy, and spectral accuracy. The system can be used for color, VIS, UV, UV-excitation-VIS-emission, and NIR imaging. The system was designed for use in a cultural heritage institution's photographic studio.

Archiving 2018 Conference Reception

17:45 – 20:15

See ticket in your registration packet for location details.

Join colleagues for this year's conference reception, just a short walk from NARA.

THURSDAY APRIL 19, 2018

Thursday Keynote and IS&T AWARDS

Session Chair: Don Williams, Image Science Associates (US)

9:00 – 10:10

9:00 **Opening Remarks and Presentation of IS&T Service Awards**

9:10 **Enhancing Access to Collections, Partnering with the Public and Enriching the Museum and Archives Fields: The Robert F. Smith Fund at the National Museum of African American History and Culture,** *Doretha Williams, National Museum of African American History and Culture (US)*

This talk discusses the implications and implementation of the Robert F. Smith Fund at the National Museum of African American History and Culture (NMAAHC). The Fund makes historical collections accessible through digitization, public programming and interaction, and support of educational development in the museum and archives fields. Through the community curation project, professional curation program, interns and fellowships opportunities, and the Explore Your Family History Center, the Smith Fund serves as a major public outreach component for NMAAHC.

2-minute Interactive Paper Previews

Session Chair: Don Williams, Image Science Associates (US)

10:10 – 10:30

FaceMatch: A System for Dynamic Search and Retrieval of Faces, *Dharitri Misra and Michael J. Gill, National Library of Medicine (US).* . . . **53**

Considerable progress has been made in recent years in locating and recognizing faces using advanced machine learning and computer vision techniques and a number of interactive tools are available for general use by individuals to use these technique in an ad hoc manner. However, no known, easily accessible open source framework presently exists to meet organizational needs to search for given faces against their pre-stored image sets, either in data analytics efforts or in time-critical situations, which leverages these techniques.

Consequently, at the US National Library of Medicine (NLM), we have implemented a Web based, publicly accessible system called FaceMatch (FM), which provides customized face matching services to clients through programmatic interfaces to robust face recognition software. In addition, it provides tools for use by the clients to submit requests in interactive or batch environment and visually observe the returned results.

Two key aspects of the FaceMatch system are: (a) it stores the contents of client's images in a repository through their key features instead of pixel values, avoiding potential copyright and other legal problems; (b) it assures near real-time availability of a newly ingested image for subsequent searches, eliminating perceptible delay between ingest and query, which is quite important in time-critical situation such as a natural or manmade disaster.

In addition to providing an HTML/REST-based API for clients to send requests to the FaceMatch server programmatically, the system also provides tools (integrated into a Java application called the FM Workbench), which allow users to submit requests interactively or in batch mode, and to visualize the returned results in real-time.

The FaceMatch system has been built for use with NLM's PEOPLE LOCATOR® Service, replacing an earlier, proprietary visual search system, and is available to be used by others with similar needs.

Automate. Now.

VERSTILE, CAPABLE, SIMPLE
FADGI 4-STAR COMPLIANT FOR ALL METRICS

Ready. Set. Go.

Digitize Everything
Bound or loose material, film and more

Automated
Precise, Exact, Repeateable movements with the new DT Autocolumn

New Software
Step 1 - Click Scan. Step 2 - Relax. You're finished. *COMING SOON*

Highest Resolution
Digitization from 40 to 100+ megapixel ready and PPI up to 6000

DT ATOM

SEE MORE AT
DTDCH.com/ATOM

(877) 367-8537 | info@dtdch.com

Long Term Preservation of Websites, *Alexander Herschung, startext GmbH (Germany)* **58**
While websites are of great interest for digital archives, the digital long term preservation of websites poses a huge problem. Given that websites consist of a large number of file formats they require today's hardware and software environment to work properly. PABLO is a software tool that processes websites and transforms them into a dramatically simplified form that is simple enough for digital archiving and exhaustive enough to preserve the websites content and appearance. The software allows users to browse the entire site like the original.

Provenance-Oriented Documentation of Multi-Spectral Data, *Ya-Ning Chen, Tamkang University; Simon Lin, Academia Sinica; M. James Shyu, Chinese Culture University; and Eric Yen, Academia Sinica (Taiwan)*. **60**
Nowadays scientific data can be generated and collected very easily with advanced instruments and facilities. In recent years, scientific data or datasets have emerged as an additional important source of scholarly output. Data-centric research needs data as a catalyst to inspire new research by repurposing or combining existing research data. Currently, most primary scientific data lack good documentation and management for future reuse, and are usually locked in personal archives as a part of dark data/archives that are in danger of being lost. Although data documentation is tedious and labor-intensive, data documentation is still cheaper than data reproduction or recollection.

Bridging Multi-Light & Multi-Spectral Images to Study, Preserve and Disseminate Archival Documents, *Bruno Vandermeulen, Hendrik Hameeuw, Lieve Watteeuw, Luc Van Gool, and Marc Proesmans; KU Leuven (Belgium)* **64**
Producing relevant photographic records in collections with unique and often fragile heritage objects is a serious challenge. Combining both visual and analytic information into such images is a major asset. Over the last decade the development of the Leuven University Portable Light Dome (PLD) has produced numerous such complex datasets. Its outcome enables different visualizations and analyses on one and the same multi-light and/or multi-spectral dataset. These interactive images support various types of research questions and contain many facets of information (reflectance characteristics, surface orientations, multi-spectral). Compared to normal photography they contain much more layered information on the archived objects. Compared to other multi-light reflectance imaging solutions such as RTI, the imaging protocol of the PLD system is able to disseminate its outcome in a multi-modal manner, beyond the visual aspects of the imaged surface.

ECHOES—Cooperation Across Heritage Disciplines, Institutes and Borders, *Ariela Netiv and Walther Hasselo, Heritage Leiden (the Netherlands)* **70**
ECHOES is an acronym for Empowering Communities with a Heritage Open EcoSystem.

That is exactly what the result of the project should be: an open system which can be used by (groups of) heritage organizations to connect data on very diverse types of heritage objects and information objects related to heritage.

Archiving Information Workflows, *Marie Vans, HP Inc., and Steven Simske, Colorado State University (US)* **75**
Understanding the context that accompanies content is an important aspect of archiving documents. In a twist on that idea, we present a new approach to archiving information workflows that includes context. Using language translation as an example of context-shifting relevancy for different tasks in an object workflow, we show how device options can determine the language presented to the user. This context-shifting behavior is important information that should be archived along with the content in a workflow.

Rare Items, Precious Time: Devising an Efficient Workflow to Digitize Nineteenth Century Cased Photographs, *Amy McCrory, The Ohio State University Libraries (US)* **77**
This paper describes an effort to standardize digitization workflows for a large collection of nineteenth century cased photographs. Using what was learned during a project to digitize selected pieces for an exhibit, guidelines and diagrams to be used in digitizing the rest of the collection were created. The process had to take into account many factors, including the curator's multiple requirements for digitization; coordination with conservation treatment of the objects; the diversity of the materials in the collection; and processes involved in transporting the items between buildings on opposite sides of a large university campus. The guidelines were written with the goal of making the process more efficient so digitization of the rest of the collection could proceed at scale, as well as minimizing the time the photos would be outside of storage. They are presented here as a model for an organized approach to digitizing a substantial number of specialized objects.

Interactive Paper (Poster) Session

10:30 – 11:20
with Coffee Break and Exhibits Open

Data Analysis

Session Chair: Erik Landsberg, Cultural Heritage Digitization Consulting (US)
11:20 – 12:35

11:20 **OCR: Unleash the Hidden Information,** *Anssi Jääskeläinen and Liisa Uosukainen, South-Eastern Finland University of Applied Sciences/Digitalia Research Center (Finland)* **83**
Most of us, even though it is not very rational, commonly take pictures of texts. In a conference it is very unlikely not to see participants taking pictures of presentation slides. Similarly, national archives scan documents without doing an OCR (Optical Character Recognition). Resulting image, in spite of its resolution, quality or file format is not searchable by its content. Unless someone types in a large amount of metadata according to Dublin Core for example. While this is an acceptable behavior in an archival world, an average people is willing to fill the maximum of five fields. Therefore a clear need for an easy and most importantly a free way to get pictures, scanned documents etc. to be fully searchable is a mandatory need.

A Digitalia research center has been working on to create an effective workflow that automatically analyzes the document content, generates OCR information as well as gets the most relevant keywords for the content. Furthermore, the workflow produces an archival graded PDF/A file if requested by the user. This workflow has been fully integrated into our Citizen Archive solution to handle everything automatically in the background. With this sophisticated solution usability, findability as well as reusability of the preserved content will be greatly increased. In short this equals better archival user experience and less manual work to be done for both the archivist and the end user.

1:45 **Research on Applying Speech Recognition for Audio-Visual Records at the National Archives of Korea,** *Jae-Pyeong Kim, Yong-Min Shin, Sang-Kook Kim, National Archives of Korea (South Korea)* **88**

Speech recognition technology can help searching and understanding the contents of audio-visual records that archives hold. But old video records sometimes do not guarantee good recognition results due to low signal quality or lack of vocabulary used at that time.

This paper shows actual experimental results and trials to enhance the accuracy using speech recognition toolkit based on deep learning, by training with relevant corpus data for video records in the 1950s and 1970s.

This paper also proposes a strategy for records management applications, considering of accuracies and service purposes for the future.

12:10 **IBRelight: An Image-Based 3D Renderer for Cultural Heritage,** *Michael Tetzlaff and Gary Meyer, University of Minnesota, and Alex Kautz, University of Rochester (US)* **93**

IBRelight is an interactive image based rendering program that allows archivists to create realistic pictures of shiny, inhomogeneous, and three-dimensional cultural heritage artifacts from flash photographs of those objects. The software provides an easy to use interface that has features similar to those provided by existing computer graphic rendering packages, but it is built on previously developed technology that can generate new images from novel viewpoints while relighting the object using point light and environmental lighting setups. Because the rendered image is created directly from the original photographs, it retains the visual fidelity of those photos, and the rendered tristimulus values can be interpreted using color management information archived with the flash photographs.

12:35 – 14:00 Lunch Break

Workflow & Quality

Session Chair: Jeanine Nault, Library of Congress (US)

14:00 – 15:15

14:00 **Dos and Don'ts for Digitisation Workflows,** *Steffen Hankiewicz, intranda GmbH (Germany)* **99**

Digitisation projects are generally complex and usually turn out to be more time-consuming than initially expected. The exact nature of the desired results should be determined well before the project starts – partly because whoever is funding the project will usually have made this a condition of support, but also because many similar projects will already have been carried out and these can be used as a guide. Yet many digitisation projects are launched without harnessing the available synergies. New software is implemented. Data formats are redesigned. In some cases, the entire system of project organisation is reinvented and tailored to meet the demands of a single project. Why?

This paper describes some of the typical pitfalls associated with digitisation project workflows and explains how even very large projects can be managed without reinventing the wheel.

14:25 **Establishing a Roadmap for Scene-Referred Raw Imaging Workflow,** *Scott Geffert, The Metropolitan Museum of Art (US)* **105**

In an effort to address current limitations, this paper documents known raw imaging shortcomings, and puts forth a roadmap for a standardized scene-referred raw workflow. In addition to traditional 2D artwork reproduction, RTI, multispectral, photogrammetry and other computational methods benefit from direct access to raw data. Due to a lack of standardization, users struggle with quality and repeatability. A key to improving raw imaging workflow is to define a raw scene-referred rendering state that can be embedded, archived and accessible to raw processors and computational imaging software as a consistent baseline. Many of the required user controls and readouts are outlined in ISO 17321-3 and ISO 19262,3,4 but until the industry fully embraces standardized raw, computation, workflow and archiving are compromised.

14:50 **Quality Assurance—Visual Inspection of Digitized Images,** *Martina Hoffmann, National Library of the Netherlands (the Netherlands)* **110**

In (mass) digitization it is common practice to work according to guidelines such as FADGI or Metamorfoze, to measure and monitor daily targets. Therefore it is common to implement some kind of quality assurance to assess if target values are met. There also is a huge field of digital preservation to ensure data can be stored appropriately. However, what about visual inspection of all those produced digital images? Why would we need to inspect the images? Visual inspection costs time, so how can we make it efficient enough, while maintaining high quality standards? How can we create a workflow for it? In this paper I will try to answer those questions based on experience from past years with a successfully implemented QA-workflow in the Netherlands.

Behind-the-Scenes Tours

15:45 – 18:00

Separate registration required. Visit the registration desk for details.

IS&T would like to thank the following Behind-the-Scenes Tour providers

Folger Shakespeare Library

Library of Congress: American Folklife Center, Veterans History Project, and Preservation Directorate

National Archives and Records Administration: Innovation Hub

National Gallery of Art: Division of Imaging & Visual Services

National Museum of African American History and Culture: Media Lab and Oral History Studio

National Museum of Natural History: DPO Herbarium Digitization Project

Smithsonian Insitute: Museum Support Center: Anthropology Digital Imaging Studio

Smithsonian Insitution Archives: Mass Digitization Project

Special thanks to NARA and its staff for their support of Archiving 2018.

FRIDAY APRIL 20, 2018

Closing Keynote

Session Chair: Don Williams, Image Science Associates (US)

9:00 – 10:00

9:00 **Obsolete Media Award for Best Interactive Paper Presented**

9:05 **Sound Preservation: Not Fast-Enough-Forward,** *Sam Brylawski, University of California, Santa Barbara (US)*

Most sound archives in the United States are relatively new, barely more than 50 years old. This talk reviews the history of institutional sound collections, assesses their current state, and considers the future of the field of acquiring, preserving, and providing access to recorded sound. The talk includes the findings of the National Recording Preservation Board's study of the state of recorded sound preservation and the resultant Library of Congress National Recording Preservation Plan, both of which were co-authored by the speaker. This overview of where we've been and where we're going is strongly colored by the personal views, priorities, and prejudices of the speaker, and his 40-plus years working with audio collections.

Databases and Data Modelling for Archiving

Session Chairs: André Kilchenmann, University of Basel (Switzerland) and Christoph Voges, Hochschule für angewandte Wissenschaft und Kunst (HAWK), and consultant (Germany)

10:00 – 14:50

10:00 **Crosswalking or Jaywalking? The Visualization of Linked Scientific and Humanities Data,** *Fenella G. France, Meghan Wilson, Chris Bolser, and Alberto Camagnolo, Library of Congress (US)* **115**

A critical aspect of shared data is using an easily accessible interface that is interoperable across a wide range of heritage institutions. An innovative approach to heritage science, where data is generated about the materiality of heritage materials, is linking this data back to a visual rendering of the heritage material to begin a process of linked data and integration between science and humanities. Using the International Image Interoperability Framework (IIIF), the shared canvas data model is being expanded for integrating linked scientific analyses to this digital surrogate. There are challenges with this approach for spectral imaging data due to the additional required layers of metadata in the spectral, spatial and temporal modes, which need to be consistent, and persistent, across sets of canvases.

10:25 – 11:00 AM Coffee Break and Interactive Paper Discussions

11:00 **A Complex Database for Documentation of Cuneiform Tablet Collection Enabling Cross-Domain Queries,** *Jaroslav Valach, The Czech Academy of Sciences; Petra Štefcová, National Museum; and Petr Zemánek, Charles University (Czech Republic)* **120**

The paper introduces multi-domain database for documentation of Prague's collection of cuneiform tablets. The complexity of documentation of the individual tablets is the most important innovation the database represents. It allows scholars to study the tablets in previously unachievable complexity of relations and context. Open formats and strict quality of data standards support reasonable hope of avoiding 'digital obsolescence' frequently observed in digitization projects.

11:25 **Preservation Data Modeling for Systems Interoperability: The Single SIP Model in the Bayou City DAMS,** *Bethany Scott and Andrew Weidner, University of Houston Libraries (US)* **124**

The University of Houston (UH) Libraries made an institutional commitment in late 2015 to migrate the data for its digitized cultural heritage collections to open source systems for preservation and access: Samvera, Archivematica, and ArchivesSpace. In order to ensure that preservation objects can be uniquely identified in Archivematica and referenced/accessed through the other systems, the UH Libraries implementation team has developed a "single SIP" data model in which a digital object's files and metadata are packaged individually prior to Archivematica ingest. The single SIP model provides flexibility in file management, avoids overloading Archivematica's processing capacity, and allows for direct persistent links from ArchivesSpace and Samvera to the preservation objects in Archivematica storage.

11:50 **Bring All Together—An Approach of a Multimedia Keep-Alive Archive,** *André Kilchenmann and Lukas Rosenthaler, University of Basel (Switzerland)* **129**

Research on moving images usually presents difficulties because the dynamic medium is not so easy to grasp. Existing software solutions facilitate the task, but are often limited to the medium of audio or video. But in our field—the humanities—we have a lot of various disciplines, each with its specific resource objects like photograph, text, video and audio, but also geographic information data, 3-d models or reflectance transformation imaging (RTI).

At our lab, we are developing one virtual research environment with the approach to bring all these different fields and their multimedia content together. The development includes a web-based user interface, a media (stream) server and a database arhitecture with a long-term perspective.

12:15 **Development for Audio-Visual Archiving System of The National Archives of Korea: A Case Study,** *Ji-Yong Lee and Sang-Kook Kim, The National Archives of Korea (South Korea)* **133**

The National Archives of Korea (NAK) has developed and is currently operating "Audio-visual Archiving System" to ensure easy accessibility of users to the records through digital conversion of analogue type audiovisual records and more systematic management of records using a system.

This paper covers the whole process of the system development and the actual operation of the system by describing the actual cases of records management such as building DB through digital conversion of audio-visual records, and user' s search and utilization.

Specifically, this paper contains the background of the development of the audiovisual records management system, the process of development, the whole process of the system, introduction of major management functions, introduction of automation functions for efficient work processing, linkage with existing systems and migration.

It is hoped that it will be used as a case study in the course of introduction and operation of audiovisual record management system in the future by presenting effect analysis obtained by operating the audiovisual records management system and suggesting constructive directions for management and utilization of audiovisual records.

12:40 – 14:00 Lunch Break

14:00 **Architecture, Design & Engineering—Archiving Digital Assets: Past, Present and Future,** *Kit Arrington[1], Aliza Leventhal[2], and Kate Murray[1]; [1]Library of Congress, and [2]Sasaki Associates (US)* **139**

For decades, design in the worlds of architecture, design and engineering have been digital and the software tools to support the work operate under a business model of rapid change and proprietary output. This paper reports on the outcome of a two day Summit held at the Library of Congress in November 2017 (Designing the Future Landscape: Digital Architecture, Design & Engineering Assets) bringing together creators, archivists, researchers, project managers, and standards and guidelines developers to illuminate the issues and challenges for preserving and accessing this work product, to explore new research possibilities created by design as data, and to identify initiatives contributing to addressing issues of preservation and access. Like the event itself, this paper hopes to increase awareness of the challenges and issues, and to share and encourage actions and collaborations for preserving this material. An in-depth consolidation of the themes and issues from the Summit can be found in the report written by Aliza Leventhal for the Library of Congress released in March 2018 entitled: Designing the Future Landscape: Digital Architecture, Design & Engineering Assets.

14:25 **Setting Out on an Unknown Sea—An Extremely Flexible Metadata Model for the "Engelandvaarders" Collection (A Case Sudy),** *Martijn van der Kaaij, Heron Information Management LLP (the Netherlands)* **143**

To address the very diverse and still developing requirements of maintaining and managing a growing collection of data on "Engelandvaarders" (people who escaped from the occupied Netherlands to England during the Second World War to continue the fight against the Germans), a flexible data model was proposed, built on semantic triples.

This approach was expected to result in a) an enduring ability to deal with new categories of resources b) a very significant reduction – after initial development - of the need for work on database interfaces, both for data entry and for data viewing and c) creation of a portable, platform independent and application independent dataset. These results were achieved, and in addition we discovered that the semantic approach notably improved communication on the metadata requirements within a varied group of stakeholders, volunteers and developers.

Finally, visualization benefits were expected, but the actual results surpassed those expectations.

Multispectral & 3D II

Session Chair: Roy Berns, Rochester Institute of Technology (US)

14:50 – 17:00

14:50 **Multispectral Imaging for Scientific Analysis and Preservation of Cultural Heritage Materials,** *Meghan Wilson, Fenella France, and Chris Bolser, Library of Congress (US)* **147**

Multispectral imaging is a digital imaging technique that adds depth to understanding cultural heritage collections. When adhering to standards and best practices it can afford a scientific analysis with commensurate integrity. The Library of Congress was one of the first institutions to implement this technology in their primary workflow as a standard for examination and preservation of its collection items. The Preservation Research and Testing Division (PRTD) has spent the past decade focusing on the development of standards and procedures for this imaging technology while also expanding its applications. Additionally, the Library of Congress has taken initiative in adapting their rigorous methodologies for practical integration of spectral imaging at other institutions. This technique expands the traditional concept of an image, while retaining the precision required for accuracy of reproducibility.

15:15 – 15:40 Coffee Break

15:40 **From the Inside Out: Practical Application of 3D Imaging Techniques for Art Conservation,** *Scott Geffert, Daniel Hausdorf, Joseph Coscia, Oi-cheong Lee, Dahee Han, Wilson Santiago, Frederick Sager, Matthew Cumbie, and Christina Hagelskamp, The Metropolitan Museum of Art (US)* **151**

One of the centerpieces of Chinese sculpture in the Asian Art collection at The Metropolitan Museum of Art is an early seventh-century seated Buddha (19.186). The life-size image was executed in jiā zhù 夾紵, or dry lacquer, a technique of layering woven textile saturated with Asian lacquer to model hollow three-dimensional objects. From 2016-2017, The Met's Buddha was examined and treated in the Department of Objects Conservation in preparation for the exhibition "Secrets of the Lacquer Buddha" at the Freer | Sackler Galleries in Washington, D.C. (December 9, 2017 to June 10, 2018). The exhibition brought together for the first time the only three known sixth- and seventh-century, life-size Chinese lacquer Buddha sculptures from The Metropolitan Museum, the Walters Art Museum in Baltimore, and the Freer Gallery of Art. Working in close collaboration, Met conservators, mount makers, and members of the Imaging Department designed an elaborate carbon fiber internal support for the Buddha, using state-of-the-art 3D scanning and milling technologies to safeguard this delicate work during transport and display. This paper documents the entire project from initial imaging to the successful fabrication of the required support structure.

16:05 **Integrating Optical Imaging of Mummy Mask Cartonnage,** *Michael B. Toth[1,2], Cynthia A. Toth[3], William Christens-Barry[1,4], Sina Farsiu[3], Guorong Li[3], Adam Gibson[1], and Melissa Terras[1]; [1]University College London (UK), [2]R.B. Toth Associates (US), [3]Duke University (US), and [4]Equipoise Imaging (US)* **157**

This rapid development and testing project captured data from multiple digital imaging techniques to try to see texts in papyrus mâché mummy mask cartonnage layers. Prior studies by other scholars destroyed the masks to access the papyri, denying future researcher access to the primary historical artefacts. This international, multidisciplinary project assessed the feasibility of integrating non-destructive digital imaging technologies to make texts visible in images of papyrus in mummy mask cartonnages for open research and analysis. The team used both optical multispectral imaging and optical coherence tomography at Duke University to try to detect the presence of text and offer scientifically valid approaches for documenting the initial state of objects and their production for future research and analysis without their destruction.

16:30 **High-Resolution Multispectral Imaging and Analysis Systems for the Very-Long-Term Monitoring of Photographs, Paintings, Fabrics, Documents, Books, and Other Works of Artistic and Historic Importance,** *Ken Boydston, MegaVision, Inc., (US); Henry Wilhelm, Wilhelm Imaging Research, Inc., (US); John McElhone, Canadian Photography Institute of the National Gallery of Canada (Canada); and Richard Adams, Ryerson University (Canada)* **163**

Development of new multispectral imaging and image-change analysis systems allows high-resolution, full-area, non-destructive, and zero-contact monitoring (without the necessity of removing works framed under glass

or plastic sheets from their frames) of photographs, paintings, fabrics, documents, books, and other works of artistic and historic importance with very large data sets consisting of ten thousand or more discrete colorimetric data points for the short-term and long-term monitoring of full-tonal-scale colorimetric changes (including in the UV and IR regions) in a fully-time-integrated fashion that might take place in the full image area and in the support material (recto and verso) over time. Irregularities in image deterioration and/or staining brought about by localized variations with the materials and "processing" employed in the creation of the work, the inevitably non-uniform contact with mounting, framing, and storage materials over time, and the effects of exposure to non-uniform lighting, environmental and "micro-climate" temperature and relative humidity conditions, can be assessed and compared over long periods of time in all areas of an image—including within very small image details. This paper attempts to consider the formidable technical challenges of very-long-term monitoring in the context of the now more than 2,000-year-old Dead Sea Scrolls in Israel, and the ongoing programs to systematically multispectrally-image, monitor, and preserve the delicate parchment scrolls and scroll fragments. Strategies that might be able to accomplish such very-long-term monitoring goals are suggested.

16:55 **Closing Remarks**

AUTHOR INDEX

IS&T CORPORATE MEMBERS

SUSTAINING CORPORATE MEMBERS

Adobe

hp

Qualcomm

SAMSUNG

xerox

SUPPORTING CORPORATE MEMBERS

FUJIFILM

KONICA MINOLTA

Lexmark™

DONOR CORPORATE MEMBERS

CLARIANT

DxOMARK IMAGE LABS

JPO JAPAN PATENT OFFICE

RICOH

Felix Schoeller Group
Best Performing Papers. Worldwide.

tRek®

Society for Imaging Science and Technology
7003 Kilworth Lane
Springfield, VA 22151 USA
703/642-9090; 703/642-9094 (fax)

https://doi.org/10.2352/issn.2168-3204.2018.1.0.2

New Techniques for the Digitization of Art Historical Photographic Archives - the Case of the Cini Foundation in Venice

Benoit Seguin, Lisandra Costiner, Isabella di Lenardo, Frédéric Kaplan; DHLAB, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland.

Abstract

Numerous libraries and museums hold large art historical photographic collections, numbering millions of images. Because of their non-standard format, these collections pose special challenges for digitization. This paper address these difficulties by proposing new techniques developed for the digitization of the photographic archive of the Cini Foundation. This included the creation of a custom-built circular, rotating scanner. The resulting digital images were then automatically indexed, while artificial intelligence techniques were employed in information extraction. Combined, these tools vastly sped processes which were traditionally undertaken manually, paving the way for new ways of exploring the collections.

Introduction

Museums and libraries hold large photographic archives to aid in art historical study, record changes in attribution and conditions of paintings, as well as document fragile architectural and monumental artwork. The Frick Museum in New York alone holds 1.2 million such images, while together, the fourteen largest photographic libraries house an estimated 31 million images [1]. In recent years, numerous initiatives have been launched to digitize this heritage, with the aim of ensuring preservation and fostering access [2]. Despite such efforts, problems persist in finding suitable methods for digitizing extensive collections of non-standard material both speedily and accurately. At the same time, there is a question of how to present the newly digitized information in a manner that would both aid current research and foster new scholarship.

This paper focuses on the case study of the digitization of the photographic archive of the Cini Foundation in Venice. It presents new technologies developed for the digitization, indexation and search of such images, presenting one approach of solving this complex problem. This research is part of the Replica project – a joint collaboration between the Cini Foundation and the DHLAB of the École Polytechnique Fédérale de Lausanne (EPFL).

Located on the island of San Giorgio Maggiore in Venice, the Cini Foundation houses roughly one million images, an especially rich repository of Venetian artistic and visual culture. The collection is divided into two sections: (1) a library comprising photographs mounted on cardboard cards with accompanying descriptive information, arranged thematically, and (2) the archives of a number of notable art historians. The first part of the archive has a standard format, namely of images mounted on cards of the same size, while the second varies widely. It can comprise of documents of different shapes, sizes, and include texts as well as images, single or double-sided.

This paper will focus solely on the digitization of the library and its cardboard cards, comprising roughly 330,000 documents.

The techniques developed in the process, and the challenges surpassed, however, can be generalized to the entire collection. They center on: (1) how to speedily and accurately digitize atypical materials found in a photographic archive, and (2) how to produce new digital methods for the automatic processing of this data - in particular, the extraction of textual and visual information from the raw scans.

New Digitization Tools

Given that the archival material came in a variety of sizes and contained visual as well as textual information on both sides of the page, a scanner was sought which could handle such constraints. Speed was equally important in the processing of the large collection. Another requirement was that the digitized file recorded the document's precise archival location. Finally, a design was sought that would minimize the operator isolation which occurs when undertaking the prolonged repetitive task of digitization over the course of months, ensuring both a more enjoyable work experience and heightened productivity.

To tackle these challenges, new scanning technologies were explored. Industry standards such as flatbed scanners, or the conveyor-belt scanner in use by the Smithsonian Museum [3], were unable to handle the complexity of the Cini material and its double-sided nature. Instead, this photographic material demanded a specialized tool, leading to the design of a custom-made scanner by Factum Arte (Madrid, Spain) in consultation with the DHLAB of EPFL.

The scanner was devised as a table with a circular, rotating top (diameter of 2 m) which comprised four image plates (Figure 1). Document sizes of up 594 x 420 mm, or A2 format, could be accommodated. The rotating top was controlled by a precision motor with variable speed enabling uninterrupted digitization of 1 image every 4 seconds. It was operated by a team of two people, one of whom placed the images on one of the glass plates, at the same time as digitization occurred on a second plate, and as a second operator removed the scanned images from a third plate. A sensor system would calculate the position and detect when a document was placed on the glass surface. Cameras mounted above and below the table simultaneously captured the recto and verso of each document placed on the glass plates. Flash units were designed and engineered by Factum Arte to provide the lowest

level of light for the achievement of a high quality image while minimizing glare. Finally, the hardware consisted of two cameras connected to two controllers, which in turn led to a server.

***Figure 1**. The Replica 360 scanner setup (Photograph: Factum Arte, 2016 [4])*

Two people operated the machine ensuring that work could be mutually checked and problems tackled collectively. The team aspect also fostered social interaction, minimizing operator isolation, enhancing work experience during the sustained repetitive task, and ensuring steady productivity.

In preparation for the scanning, bar-codes were created to record the archival position of each document, and were affixed to the verso of each cardboard card. To speed up the workflow, an xml file was also created. This automatically named files as they were being digitized with the title of the collection and the document's position within this, ensuring that scanning progressed continuously without the need to stop and name files individually. For each document in the archive, four files were created: one for the digitization of the recto side, one for the verso, and one corresponding md5 signature file for each of these to verify authenticity and integrity. Accuracy and integrity of files could likewise be checked by comparing the automatically generated name with the bar-code.

In the case of the Cini Foundation, the design of the scanner allowed for the digitization of 1500 images per day, with the digitization of the entire photographic archive of 330,000 images completed in roughly 18 months.

Image Processing Pipeline

The scanner output resulted in high-resolution raw images (of 400 ppi and 5424 x 3616 pixels). As the scanning of the material took place on-site at the Cini Foundation in Venice, and the data was processed by the DHLAB of EPFL, Switzerland, transferring the data posed a challenge in itself. The 50 terabytes of RAW files, were converted to JPEG images of 90% quality. Only the scans of the document rectos were transmitted, reducing the size to 2.6 terabytes.

Figure 2 illustrates a typical scan of the recto side of a document. Each of the Cini scans contained the reproduced artwork pasted on a cardboard card. Metadata information about the artwork is typed at the top of the card. The scan likewise includes a black border surrounding the image, as well as a color control bar. To process the image, the relevant information had to be extracted. This included, cropping the art reproduction from the scan. Secondly, the structured metadata, written at the top of each cardboard, had to be extracted from the image and rendered into text, to enable a future textual search of this information. Because thousands of scans awaited processing, devising an automatic approach to complete these tasks was essential.

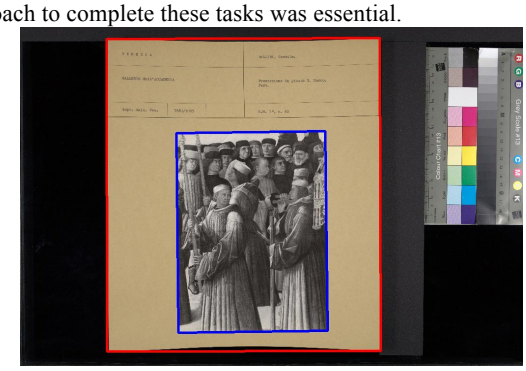

***Figure 2**. Scanned image. The red and blue rectangles denote the areas to be extracted – respectively, the cardboard card, and the art reproduction.*

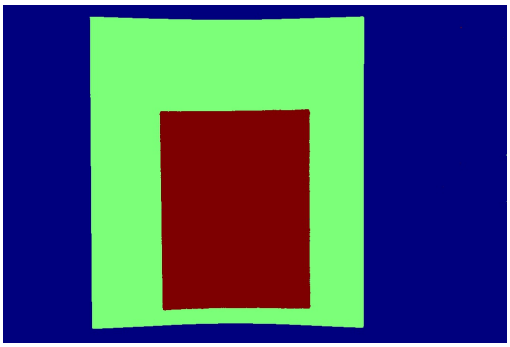

***Figure 3**. Predicted mask from the corresponding scanned image in Figure 1. The annotated training mask is extremely similar to this.*

Image segmentation

Segmenting the different areas of interest proved complicated. Within each scanned image, the cardboard could appear in different positions, and orientations, at times being rotated up to 90°. Aging and humidity also affected some documents, deforming the cardboard support so that it no longer appeared rectangular. Difficulties were compounded by inconsistent scanning practices in the early days of digitization, which likewise produced non-standard layouts with the color control bar at times overlapping the area of the cardboard. Upon these cardboard supports, the art reproductions also varied in position, shape and orientation. At times, these images filled the entire area of the cardboard obscuring the metadata. Their colors and textures likewise differed. Some monochrome photographs even resembled in color their cardboard backing. This variability posed challenges for automatic extraction, as there was no consistent digital marker across images.

Given this variety, standard techniques for image segmentation that rely on only one visual cue were inadequate. Color and texture were not sufficiently discriminant. The detection of lines which were predominantly straight and the creation of a

rectangle out of these, was the technique that appeared most promising, but it too failed in yielding perfect accuracy.

The approach which produced the best results was based upon artificial intelligence, namely deep learning and convolutional neural networks (CNN). More specifically, we used a specific form of CNN called a fully-convolutional neural network (FCN). These networks took as input an image and returned as output an image of the same size where each pixel contained a number of class probabilities. In the past, these techniques proved effective for pixel-wise predictive tasks, and they currently underlie the best performing systems for semantic and instance segmentation [5]. FCNs have recently been introduced successfully to tackle problems of document processing, especially historical document processing. The model used in the course of this project is based upon a U-net architecture [6] where most parameters derive from a pre-trained version of the Resnet-50 architecture [7]. Details of the architecture of this system and how it was trained rest outside the scope of this paper and are described elsewhere [8].

In short, to process the Cini documents, a FCN was trained to predict for each pixel whether it belonged to the background, cardboard backing, or the art reproduction. Training the system required a corpus of correctly annotated images. These were produced by taking a number of scanned images and drawing upon them with an image editing software. Here, the background was painted one solid color, the cardboard area another color, and the art reproduction area with a third color (as in Figure 3). Through this process, a user annotated 120 training images in the course of 2 hours.

The trained FCN generated an image along with the probability of each pixel of being part of the background, the art reproduction, or the cardboard. The highest probability was taken, yielding a prediction image, where each pixel was assigned to one of the three classes (Figure 3).

From this prediction image, the smallest rectangle that would enclose all the pixels of a given class was extracted. Additional cleaning, based on morphological operations, further removed small artifacts, or outlying pixels, that arose in the course of the prediction. Additional logic was used to improve the results in specific cases. For instance, at times, the FCN confused black non-textured areas of the art reproduction with the background of the scan. Imposing the rule that the art reproduction area has to be located within the space of the extracted cardboard corrected this issue.

Reading the metadata

The next step in image processing centered on extracting the textual metadata that was printed at the top of each cardboard card. Fortunately, the large majority of this metadata was typed and not handwritten, enabling optical character recognition. For this task, standard optical character recognition (OCR) libraries were tested, including the open-source Tesseract and the commercial GoogleVision [9]. The latter performed much better than the former, so GoogleVision was used in this study.

An OCR algorithm provided bounding boxes around each word and its respective transcription. These words were then clustered together into small paragraphs so that each paragraph accounted for a single metadata entry.

Next, each paragraph had to be assigned the correct metadata label (author, description, city, etc.). Here too the metadata layout of the Cini cardboard cards proved inconsistent. Although the same categories were typically relegated to roughly the same areas of the cardboard, the spaces allocated to these fields differed in height. Furthermore, the lines denoting the separation between fields were not always straight because of the bending of the cardboard, rendering this structure unreliable for establishing divisions.

To account for these inconsistencies, the assignment of metadata labels relied on the position of the extracted paragraphs. Given a layout model (which is a collection of pairs <position, label>), we found the best label for each paragraph by looking at the distances between the paragraph's centroid and its position within the layout. Multiple layout models (encoding the variability of the headers) were also tested, picking the one that best matched the extracted paragraphs (based on the sum of the distances between the paragraphs' centroids and their respective assigned position in the layout).

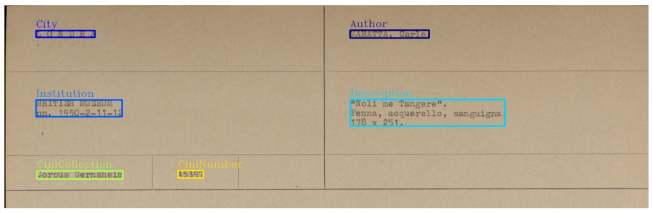

Figure 4*. Close-up of a simple metadata header, with the detected text and the corresponding metadata label assigned plotted in color on top.*

Evaluation

To test the efficiently of the processing pipeline and its chosen tools, the quality of the results was evaluated. In particular, we evaluated: (1) the extraction of the visual elements, (2) the detection of the text, (3) the label assignment, and (4) the OCR quality.

For the extraction of the cardboard cards and the art reproductions, we annotated 120 scans that were used for training the FCN (100 training + 20 validation), and then tested the performance on 150 additional examples. The metric used to evaluate the bounding rectangle and its correspondence to the ground-truth was the intersection-over-union (IoU) of these areas. As the results show (Table 1), near perfect extraction accuracy was achieved.

In terms of text detection and the evaluation of the OCR quality, we manually examined 412 random cardboard cards. For this step, the biggest problem was created by the archival practice of manually correcting the typed metadata with handwritten annotations. Unsurprisingly, in these cases, the OCR rarely transcribed the handwritten text. Two types of error were predominant in this case: ones where the cardboard omitted information in one category (i.e. where a particular field was left blank), or where information was partially missing. For instance, the OCR might miss the second line of the author field which contained additional information, or omit the handwritten annotation which was a precision and not a full correction. Results are shown in Table 2.

For the label assignment, we evaluated 337 cardboard cards where all the text elements were properly detected, corresponding to 2028 metadata fields. The only issues recorded were 5 cases of fields being merged together and 3 cases of one field being separated into two.

For the OCR quality, we manually corrected the transcription of the author and description fields (which are the most important for our project), in a set of 150 cardboard cards. The results are displayed in Table 3, and show a good accuracy for the OCR algorithm. Most of the fields are perfectly transcribed with at most a 1-character error, which makes the transcriptions good-enough for undertaking 'fuzzy searches', and allows further automatic correction of these errors with an artist name dictionary, such as the Getty's Union List of Artist Names [10]. Still, some challenges remain, including the detection and reading of handwritten annotations, and the flagging of ambiguous cases that require manual review.

	Cardboard mean-IoU	*Photograph mean-IoU*	*Photograph IoU>95%*
-Base predictions	99.3%	98.2%	96.7%
-Cleaned predictions	99.3%	98.2%	97.3%
-Additional logic	99.3%	99.0%	100.0%

***Table 1**. Results of the quality of the image extraction across 150 randomly chosen scans. The second row correspond to simple post-processing cleaning operations of the predictions (morphological operations). The last row corresponds to adding the additional logic of the photograph being inside the cardboard.*

Metadata Field	*Completely missing*	*Partially missing*	***Total***
Author	1.7%	1.7%	**3.4%**
Description	0.5%	2.4%	**2.9%**
Institution	0.2%	1.0%	**1.2%**
City	1.7%	0.7%	**2.4%**

***Table 2**. Text detection error-rate for some fields, evaluated on 412 random scans.*

Metadata Field	*Perfect*	*<=1 error*	*Perfect (normalized)*	*<=1 error (normalized)*
Author	77.3%	96.62%	83.8%	97.3%
Description	77.3%	93.92%	85.8%	96.6%

***Table 3**. Quality of the OCR transcription for 150 cardboard cards where text is detected. The average character-error-rate is 2.0% and 1.4% for the author field and the description field respectively. The "normalized" column indicates that we do not consider lowercase/uppercase errors, nor blank spaces/punctuation.*

Conclusion

This study details the digitization of the photographic archive of the Cini Foundation introducing techniques that address the challenges posed by complex archival materials. It proposes new tools for the rapid digitization of photographic archives and details digital techniques for image segmentation and information extraction. As cultural institutions have been undertaking much of this work manually, these techniques are poised to vastly speed the process of digitizing art historical collections, while at the same time facilitating access and assisting research.

The final results of the image extraction and metadata reading techniques outlined here display a high accuracy which renders this extracted information immediately usable. In fact, these images and their metadata are already being employed as the basis for a new search engine, Replica. This tool enables the search of similar images, or of details within images, further showcasing the usefulness of new digital tools in advancing art historical research [11] [12].

Acknowledgments

We would like to thank the Cini Foundation, in particular the Secretary General, Pasquale Gagliardi, and all of the staff and collaborators who oversaw and participated in the digitization process. We are also grateful to Factum Arte, and in particular to Adam Lowe and his team for the development and creation of the scanner. This research forms part of the Venice Time Machine project, generously supported by the Fondation Lombard Odier.

References

[1] T. Loos, "'Photo Archives Are Sleeping Beauties.' Pharos Is Their Prince." *New York Times*, 14 Mar. 2017, https://www.nytimes.com/2017/03/14/arts/design/art-history-digital-archive-museums-pharos.html.

[2] Pharos: The International Consortium of Photo Archives, http://pharosartresearch.org/.

[3] S. Overly, "The Smithsonian turned to conveyor belts, cameras to digitize its many artifacts." *The Washington Post,* 25 Jan. 2015.

[4] Factum Arte, "Press Release: Replica 360 Recto/Verso Scanner," 2016, http://www.factum-arte.com/pag/757/Replica-360-Recto-Verso-Scanner.

[5] K. He, et al, "Mask R-CNN", in ICCV, Venice, Italy, 2017.

[6] O. Ronneberger, et al, "U-Net: Convolutional Networks for Biomedical Image Segmentation", in MICCAI, 2015.

[7] K. He, et al, "Deep Residual Learning for Image Recognition", in CVPR, 2016.

[8] B. Seguin, S. Ares Oliveira, et al, "A generic segmentation approach for baseline detection, document layout analysis and object extraction in historical documents", 2018.

[9] GoogleVision, https://cloud.google.com/vision.

[10] B., Seguin, et al. "Extracting and Aligning Artist Names in Digitized Art Historical Archives", in Digital Humanities Conference, Mexico City, Mexico, 2018.

[11] B. Seguin, et al. “Visual Link Retrieval in a Database of Paintings”, in *ECCV Workshops*, pp. 753–767 DOI: 10.1007/978-3-319-46604-0 52, 2016.

[12] I. Di Lenardo, et al. “Visual Patterns Discovery in Large Databases of Paintings”, in Digital Humanities Conference, Krakow, Poland, 2016.

Author Biography

Benoit Seguin is a PhD Candidate at the Digital Humanities Laboratory at the École Polytechnique Fédérale de Lausanne. His research focuses on employing modern computer vision techniques on large visual databases of art, leading him to develop new algorithms for the digitization and exploration of these photographic collections. He received a Diplôme d'Ingénieur from École Polytechnique Paristech and a Master of Science in computer science from the École Polytechnique Fédérale de Lausanne.

Lisandra S. Costiner is a Post-Doctoral Fellow in the Digital Humanities Laboratory of the École Polytechnique Fédérale de Lausanne. Her doctoral dissertation (University of Oxford, 2017) focuses on Italian medieval and Renaissance visual and devotional culture. Before turning to art history, she pursued a Bachelor's degree in visual and environmental studies at Harvard University.

Isabella di Lenardo, holds a doctorate in Theories and Art History, and is a Researcher in Digital Humanities at the École Polytechnique Fédérale de Lausanne. Her interests center on the production and circulation of artistic knowledge in the XVIth-XVIIIth centuries. She collaborates with numerous institutions in the fields of digitization and preservation of Cultural Heritage.

Frédéric Kaplan holds the Digital Humanities Chair at École Polytechnique Fédérale de Lausanne and directs the Digital Humanities Lab, leading projects that combine archive digitization, information modelling and museographic design. With his team, he is developing the Venice Time Machine, a project to model the evolution and history of Venice over a thousand years. Frédéric holds a PhD in Artificial Intelligence from University Paris VI, and previously worked as researcher at Sony Computer Science Laboratory.