

Solving the Challenges of Digital Archiving

by

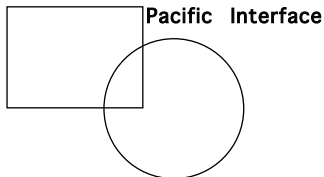
Laurin Herr

Pacific Interface, Inc.

Digital Archiving Symposium

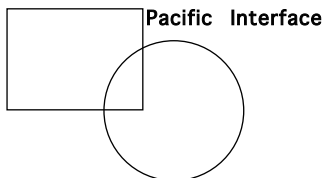
Keio University DMC

October 24, 2008



Analog vs. Digital Archiving Model

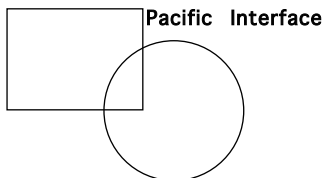
- **Longevity of traditional analog archives**
 - primarily determined by media durability
 - proper use of conservation techniques
 - preserved objects can survive periods of “benign neglect.”
- **Longevity of digital archives**
 - primarily determined by periodic “data migration”
 - physical conservation of objects not sufficient
 - “benign neglect” can be a fatal
- **Therefore, in addition to traditional archiving costs, digital archives require recurring investments to support periodic “data migration.” This makes digital archiving more expensive over the long-term and increases the need for organizational continuity and sustained funding.**
- **Historically, storage (disks and tape) has been the biggest expense. But as digital archives grow and storage prices decline over time, the ongoing costs of data management services, labor and electricity will increase as a percentage of the total budget.**



Challenges of Digital Archiving

“Immensity”

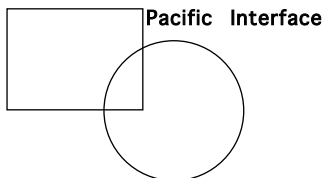
- **AMPAS estimates**
 - 2 - 10 Petabytes per digital movie in 2007 (if save everything)
- **University of California Berkeley estimates**
 - 2-3 Exabytes unique data created worldwide in 1999
 - 5 Exabytes worldwide in 2002
- **International Data Corporation estimates**
 - 161 Exabytes of data created, captured, replicated worldwide in 2006
 - 988 Exabytes worldwide in 2010
- **Enterprise Strategy Group estimates**
 - 3 Exabytes archival storage worldwide in 2005
 - 25 Exabytes to be required in 2010
- **Becoming impossible (uneconomical) to save everything.**
- **New challenge: how to decide what to discard?**



Challenges of Digital Archiving

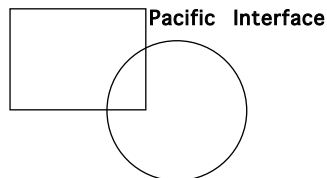
“Eternity”

- **If the mission is to preserved data “eternally,” must make some assumptions for long-term planning**
 - Digital media lifespan is not “forever”
 - Software/hardware support of a given device is not “forever”
 - Technology progresses through successive “generations”
- **Therefore, digital archive design goals must be:**
 - preservation of the data itself (not just the media)
 - continuous operation of the system overall (not just discrete hardware or software components).
- **Therefore, practical priorities for digital archivist are:**
 - Data integrity
 - Data protection
 - Data migration



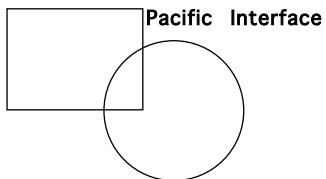
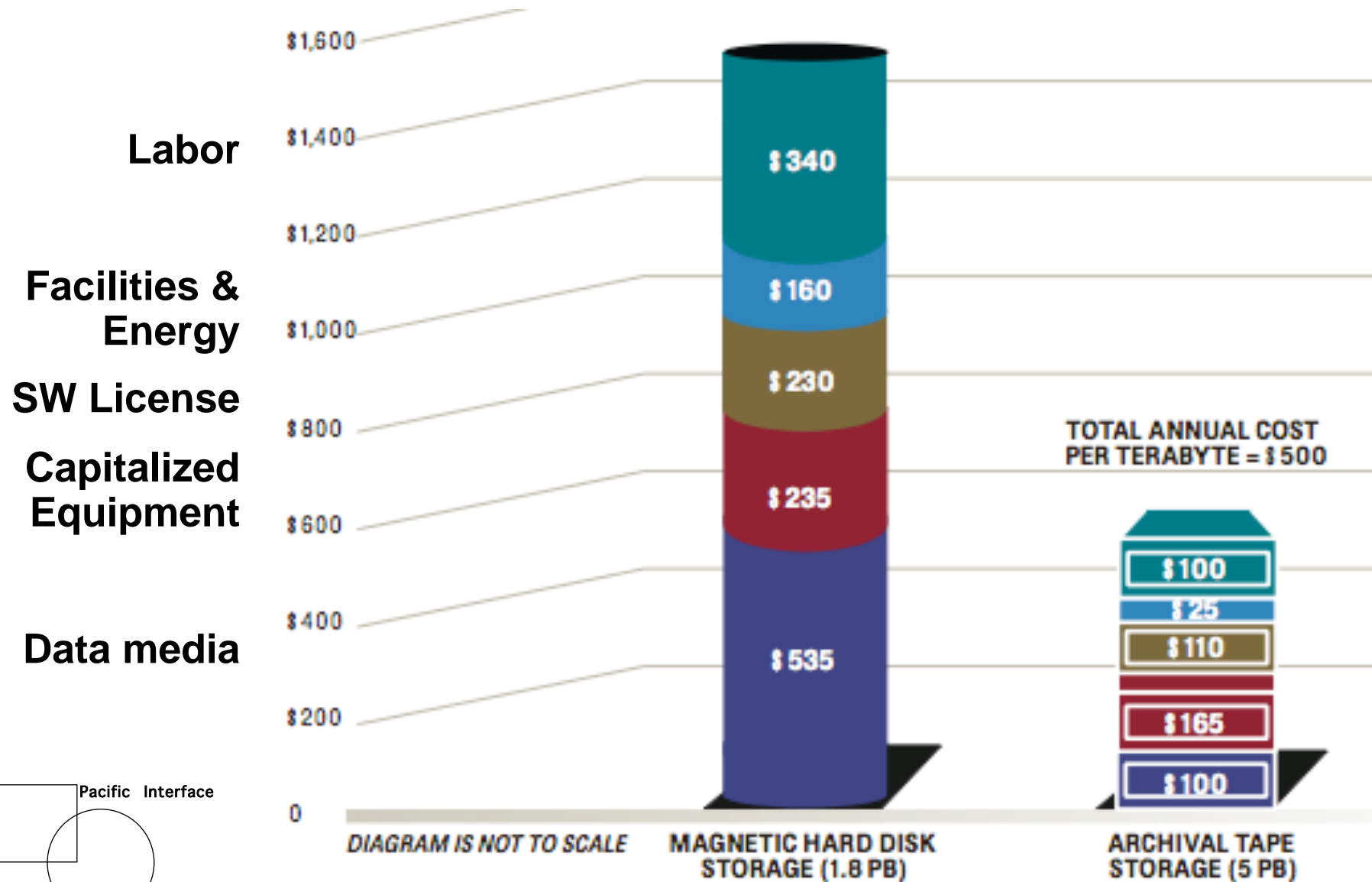
Storage Trends Are Favorable

- **Technical advances 1975 - 2005**
 - Data tape areal density up 1333x
 - Data tape transfer rate up 24x
 - Hard disk areal density up 2250x
 - Hard disk transfer rate up 21x
- **HDD shipments grow as \$/GB falls 40% per year**
 - In 1995, <200 PB of disk shipped worldwide
 - In 2000, >2000 PB of disk shipped worldwide
 - In 2005, “street price” of cheap disk was \$0.50 - \$1.00 per GB
 - In 2010, price trends point to 5¢ - 10¢ per GB
- **Magnetic tape most widely used for digital archiving**
 - Tape cost \$/GB lower than HDD
 - Energy cost of tape much lower than HDD











Challenges of Digital Archiving

“Total Cost of Ownership”



Challenges of Digital Archiving

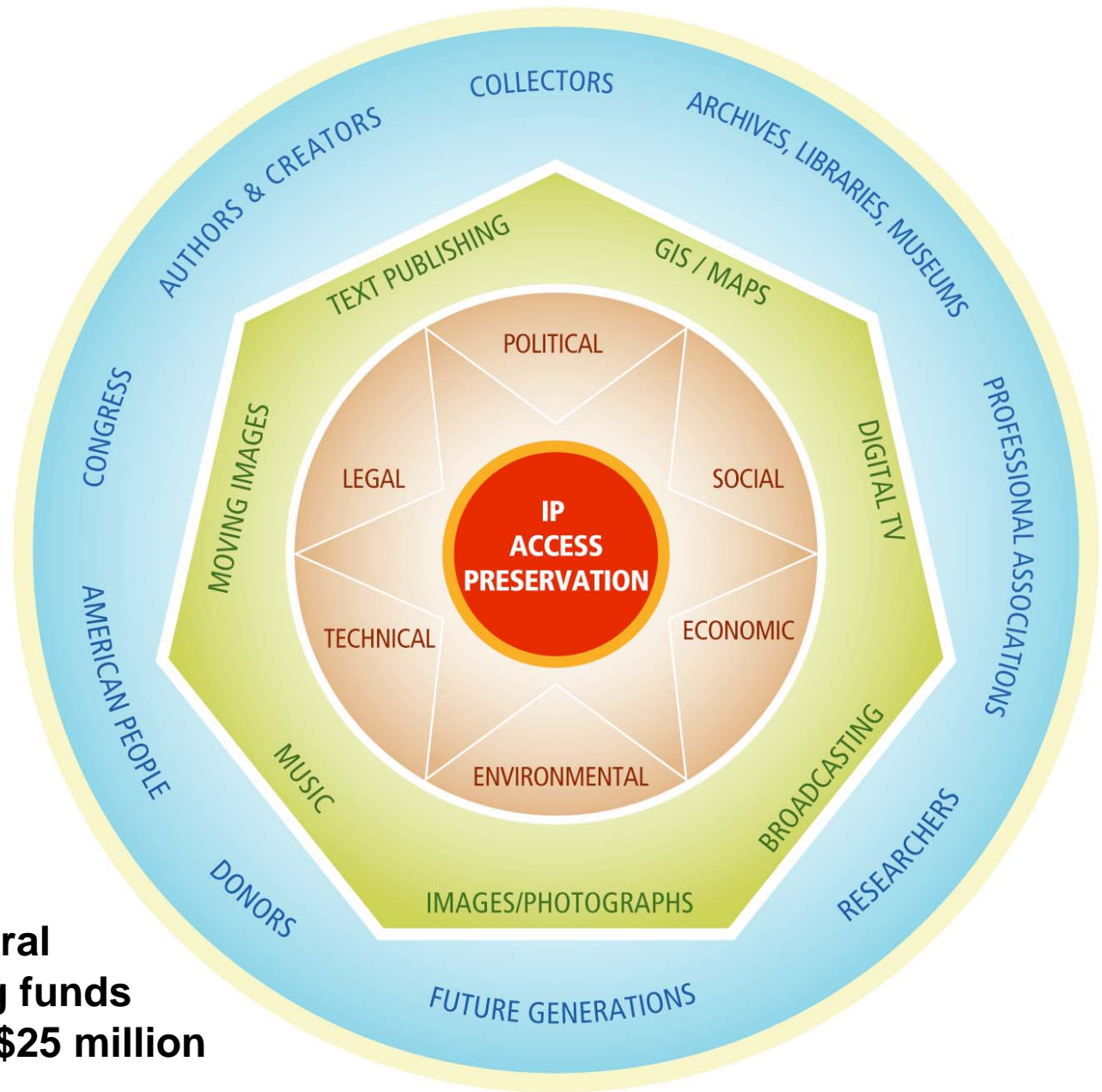
“Total System Stack”

LIFESPAN	HARDWARE	SOFTWARE
3 → 5 YEARS	 HOST COMPUTER	<ul style="list-style-type: none"> • APPLICATION SOFTWARE • OPERATING SYSTEM • DEVICE DRIVERS
5 → 10+ YEARS	 PHYSICAL INTERFACE	<ul style="list-style-type: none"> • INTERFACE FIRMWARE
3 → 5 YEARS	 MEDIA DRIVE	<ul style="list-style-type: none"> • DRIVE CONTROL FIRMWARE
.5 → 10 YEARS	 MEDIA	<ul style="list-style-type: none"> • FILE SYSTEM • DATA FILE FORMAT • PHYSICAL RECORDING FORMAT
VARIES	 TRAINED PERSONNEL 	
VARIES	 FUNDING 	

National Digital Information Infrastructure and Preservation Program **(NDIIPP)**

Collaborative Initiative
of the
US Library of Congress

Created in 2000 with Federal
budget + private matching funds
Investment to date about \$25 million



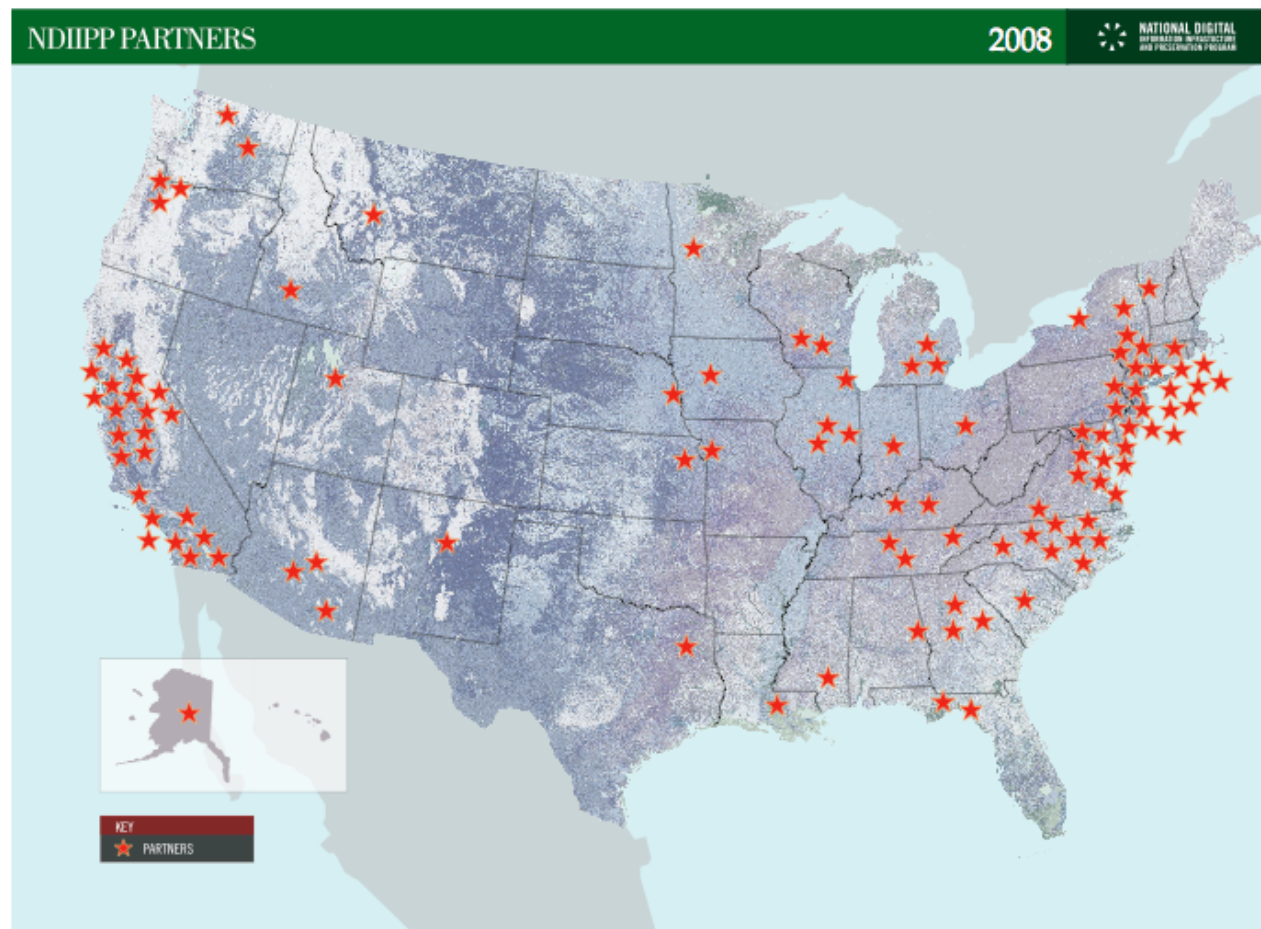
NDIIPP Lessons Learned (so far)

- **Distributed and decentralized approach required**
- **Need for research into new tools and technologies**
- **Recognition that technology is only part of the problem**
- **Intellectual Property Rights must also be cleared for long-term**
- **Need to create a balance between preservation and access**
- **Two major components of digital infrastructure**
 - A preservation network of partners
 - A technical architecture
- **Investment priorities**
 - Network of partners
 - Preservation architecture
 - Basic research

NDIIPP Network of Preservation Partners

"No one institution can tackle the challenge of digital preservation on its own."

As of 2008, the Library of Congress has over 130 partners in the digital preservation network which connects libraries, archives, universities, research centers, non-profit and for-profit organizations and professional associations.



U.S. National Archives and Records Administration (NARA)

The challenge of NARA's Electronic Records Archive

- Preserve any type of electronic record
- Created using any type of application
- On any computing platform
- From any entity in the Federal Government or any donor
- Provide discovery and delivery to anyone with an interest and legal right of access
- Now and for the life of the Republic
- **NARA's own mission is served by advocating effective policies, strategies, standards, guidance, and tools for other Federal agencies to manage electronic records in support of their governmental responsibilities**
- **NARA needs help to fulfill its own mission**





US National Archives Records Administration Electronic Records Archive Program Partners



National
Science
Foundation



San Diego
Supercomputer
Center



National Computational
Science Alliance



The Library of Congress



Army Research
Laboratory



National Agricultural Library

DIGITAL LIBRARY
FEDERATION



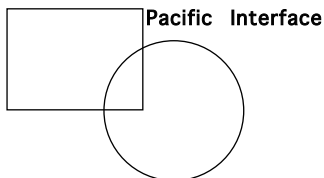
National Partnership for Advanced Computational Infrastructure



DataNet

Sustainable Digital Data Preservation and Access Network

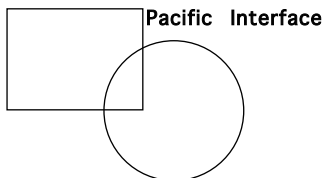
- **New NSF project to develop methods, management structures and technologies to manage the diversity, size, and complexity of current and future data sets and data streams over the long-term.**
- **Targets any information stored and accessed digitally: numeric data, text, publications, sensor streams, video, audio, algorithms, software, models and simulations, images, etc.**
- **Focus on data collections central to the scientific and engineering research and education mission of NSF.**





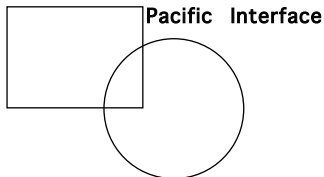
DataNet

- **Goal is to create a national and global framework for preservation and access of digital scientific and educational data collections**
 - Create new types of organizations that integrate capabilities to design, develop, operate and sustain DataNet over the long-term.
 - Facilitate research in concepts and technologies that support long-term preservation & access for complex digital objects & assets.
- **NSF plans to grant up to 5 research awards for DataNet, each up to \$4,000,000 per year for up to 5 years (\$20,000,000 per award)**
 - Two awards to be decided in late 2008
 - Three more awards to be decided in 2009



Solving the Challenges of Digital Archiving?

- **There is not just “one” answer to the digital dilemma.**
- **Storage technology is essential to digital archiving, but not sufficient by itself to solve the digital dilemma.**
- **Study the issues from every perspective**
 - Storage technology
 - System architectures
 - Organizational structures
 - Human resources
 - Institutional policies
 - Collection strategies
 - Economic models
 - Funding
- **Build community of collaborators**
- **Learn by doing smaller scale prototype archiving testbeds**
- **Start before problems become overwhelming**
- **Lead from the top**

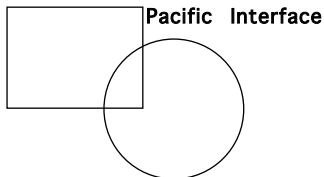


The Digital Dilemma Is Strategic!

In the Industrial Society of the 20th Century, total quality control - pioneered in the USA and perfected in Japan -- was a key to increasing manufacturers' competitiveness, internal efficiency and profitability while simultaneously offering consumers improved price/performance, better reliability and a more satisfying user experience.

In the Information Society of the 21st Century, the process of solving the digital dilemma will force data users to creatively focus on the purpose, value and meaning of the information they generate and save. This can lead to systemic improvements affecting every part of society.

On the other hand, if we can't collectively solve the digital dilemma, the next generation will inherit no meaningful historical foundation upon which they can build their future. This is a strategic challenge we must not fail.



Thank You

Laurin Herr
President, Pacific Interface Inc.
Oakland, California, USA

laurin@pacific-interface.com

