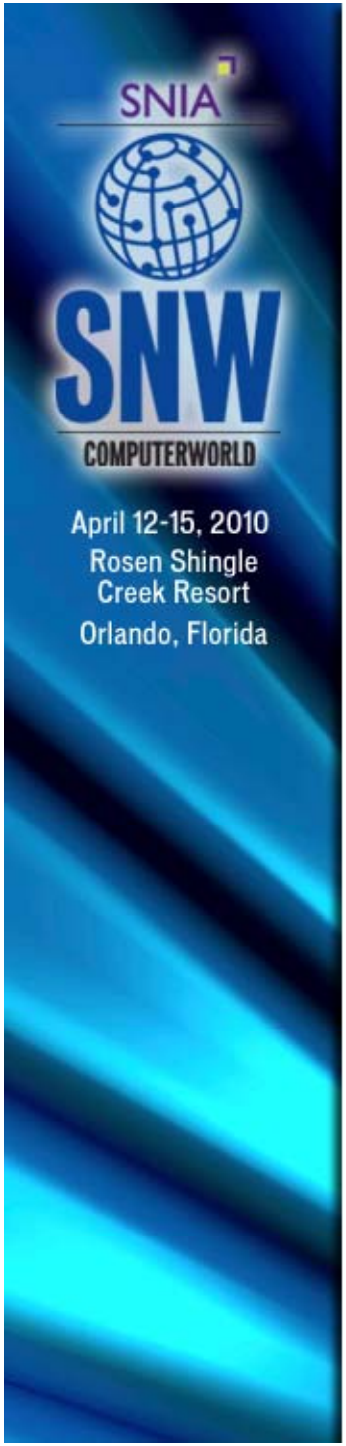

COMPUTERWORLD
SNIA
SNW
April 12-15, 2010 | Rosen Shingle Creek Resort | Orlando, Florida

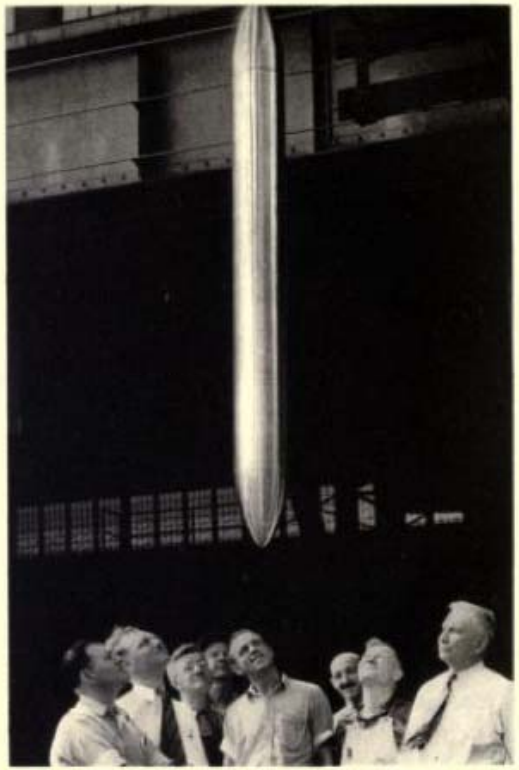


More than a Lifetime of Data and Information

Unifying Live and Archival Storage

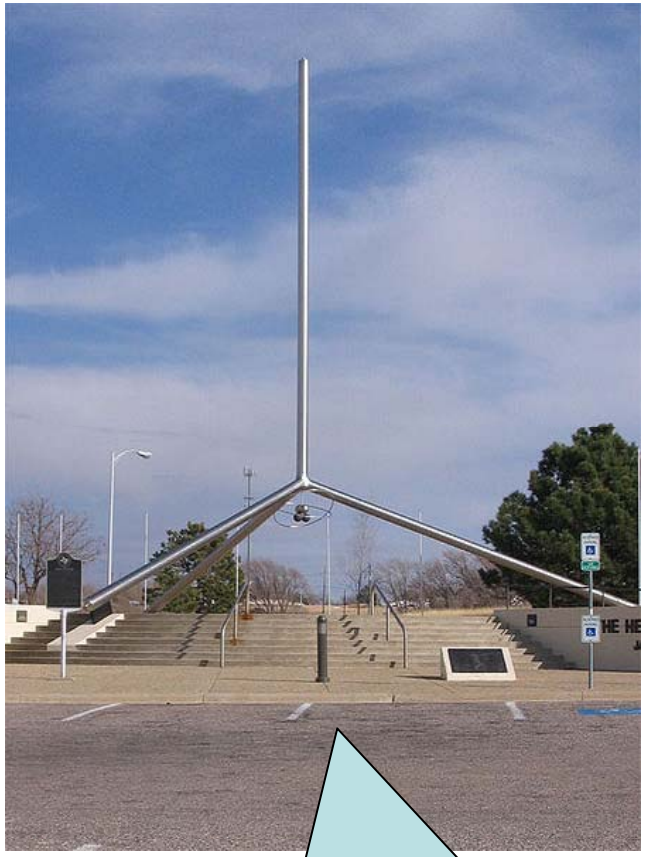
Larry Stabile
Iron Mountain Digital

Time Capsules



THE ENVELOPE FOR A MESSAGE TO THE FUTURE BEGINS ITS EPIC JOURNEY

5000 years – NY World's Fair, 1939



1000 years– Amarillo, Texas, 1968

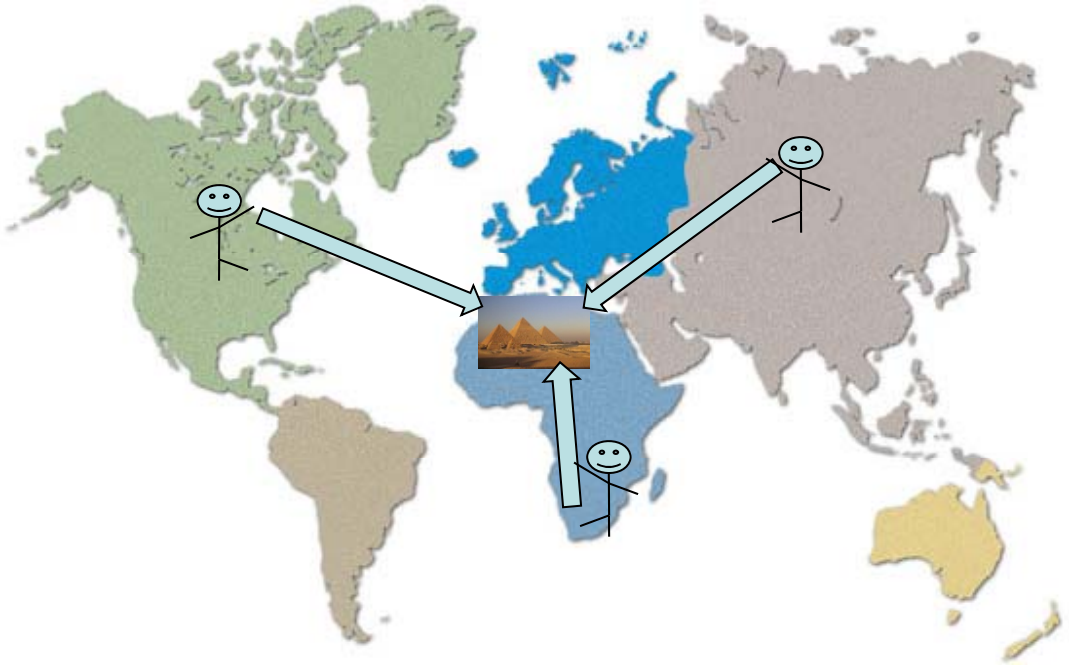
Pyramids

4500 years old

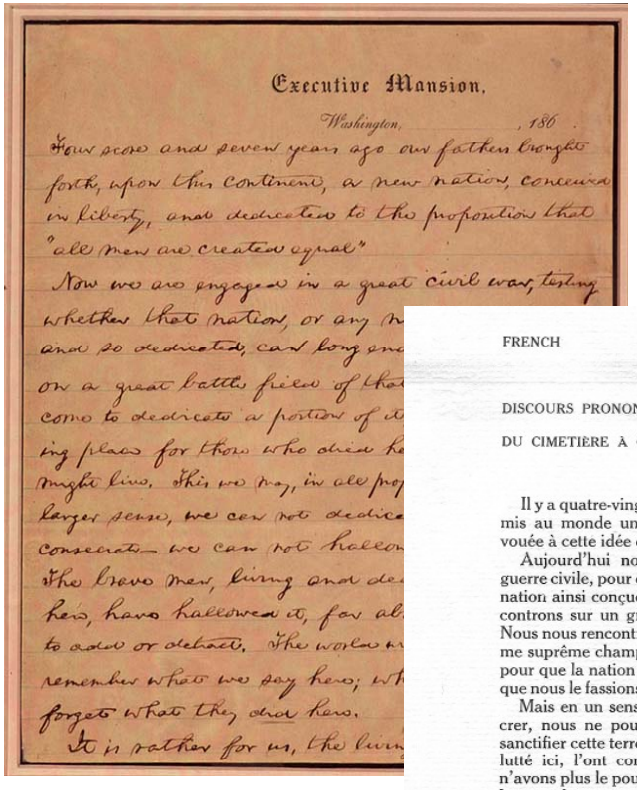
Physical Preservation of Unique Objects

Principle of Linkage

The quality of preservation increases with the number of interlinked references among the preserved items



Gettysburg Address



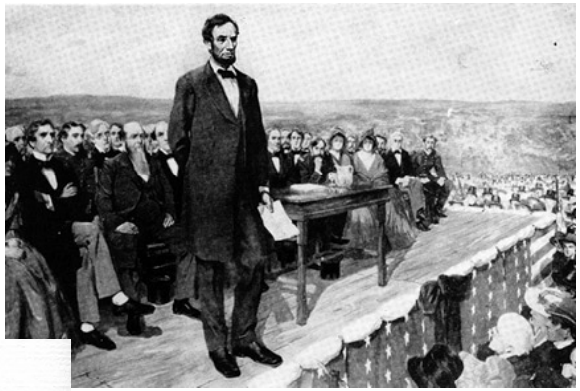
Executive Mansion,

Washington, , 186

Four score and seven years ago our fathers brought forth, upon this continent, a new nation, conceived in liberty, and dedicated to the proposition that "all men are created equal"

Now we are engaged in a great civil war, testing whether that nation, or any n and so dedicated, can long sn on a great battle field of tha come to dedicate a portion of it ing place for those who died he might live. It is, in all pro larger sense, we can not dedica consecrate we can not hallow The brave men, living and de here, have hallowed it, far al to add or detract. The world w remember what we say here; wh forget what they did here.

It is rather for us, the livi



FRENCH

DISCOURS PRONONCÉ À LA DÉDICATION

DU CIMETIÈRE À GETTYSBURG

Il y a quatre-vingt sept ans, nos pères ont, sur ce continent, mis au monde une nouvelle nation, conçue en liberté et vouée à cette idée que tous les hommes naissent égaux.

Aujourd'hui nous sommes engagés dans une grande guerre civile, pour déterminer si cette nation — ou toute autre nation ainsi conçue et dédiée — peut durer. Nous nous rencontrons sur un grand champ de bataille de cette guerre. Nous nous rencontrons pour en consacrer une parcelle, comme suprême champ de repos, à ceux qui ont donné leur vie pour que la nation puisse vivre. Il est convenable, il est juste que nous le fassions.

Mais en un sens plus large, nous ne pouvons pas consacrer, nous ne pouvons pas dédier, nous ne pouvons pas sanctifier cette terre. Tous les héros, vivants et morts, qui ont lutté ici, l'ont consacrée de manière si haute que nous n'avons plus le pouvoir d'y rien ajouter, ni d'en rien enlever. Le monde remarquera peu ce que nous disons ici et il ne s'en souviendra guère, mais il n'oubliera jamais ce que des braves ont fait en ce lieu. C'est plutôt à nous, les vivants, d'être voués à la tâche encore inachevée qu'ils ont jusqu'ici si noblement accomplie. C'est plutôt à nous d'être dédiés à la grande tâche qui nous reste — afin que ces morts vénérés nous inspirent un dévouement accru pour la cause qui leur a fait combler la mesure du dévouement — afin que nous soyons fermement résolus à ce que ces morts ne soient pas morts en vain; afin que cette nation, devant Dieu, renaisse à la liberté — et afin que le gouvernement du peuple, par le peuple, pour le peuple, ne soit pas effacé de cette terre.

ABRAHAM LINCOLN.

le 19. novembre 1863.

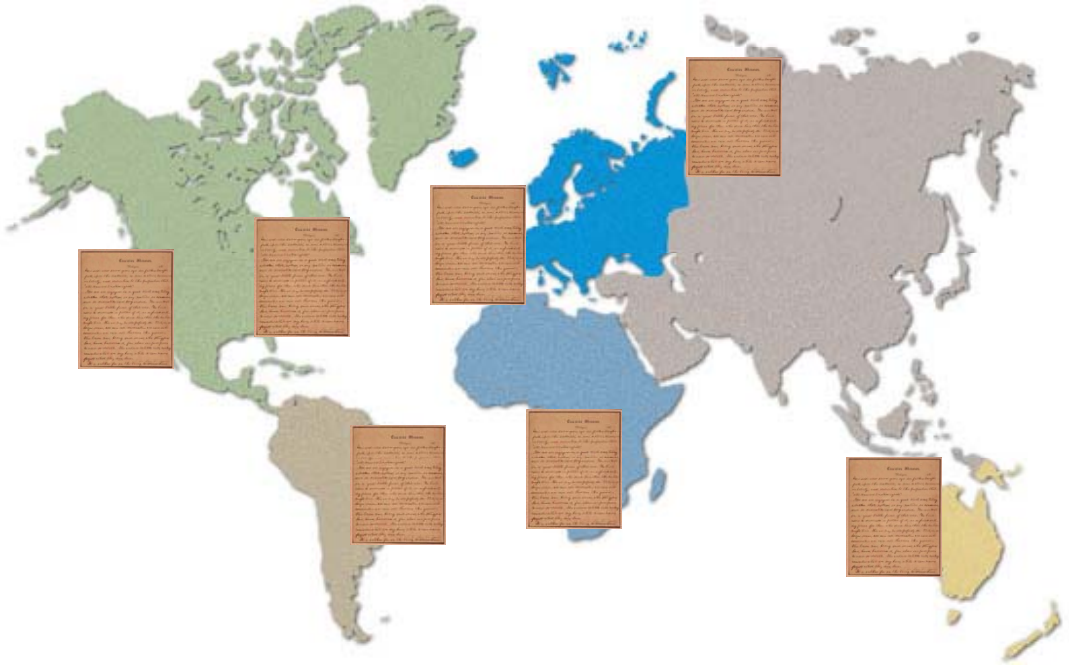
Version par André Maurois

French Translation

On the wall of the Lincoln Memorial

Principle of Redundancy

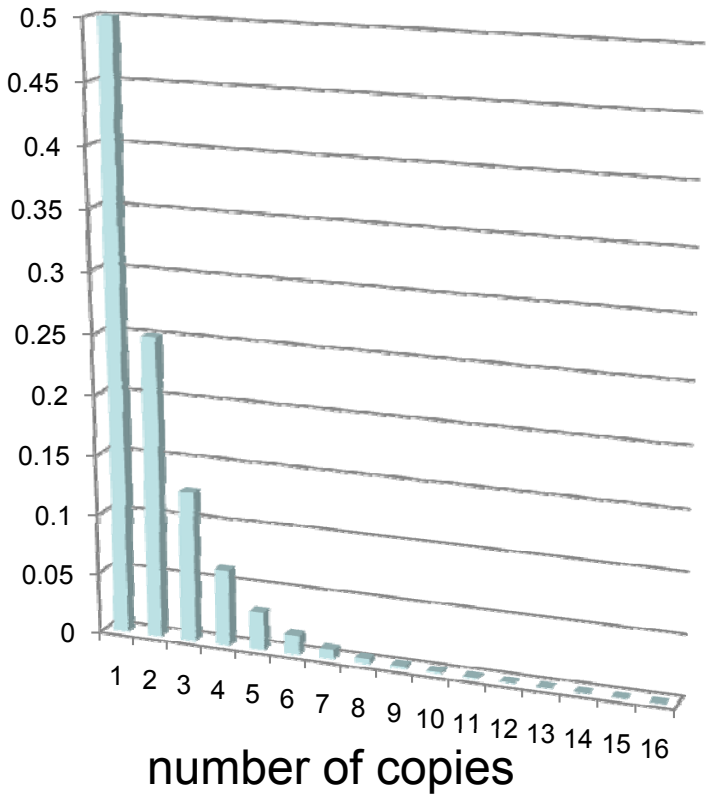
The quality of preservation increases with the number of copies of a given item and their physical dispersion.

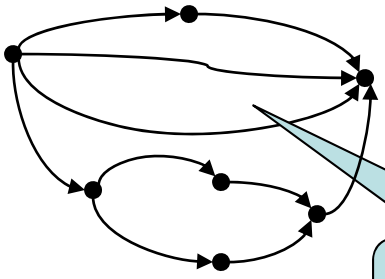


Redundancy Decreases Loss Probability Quickly

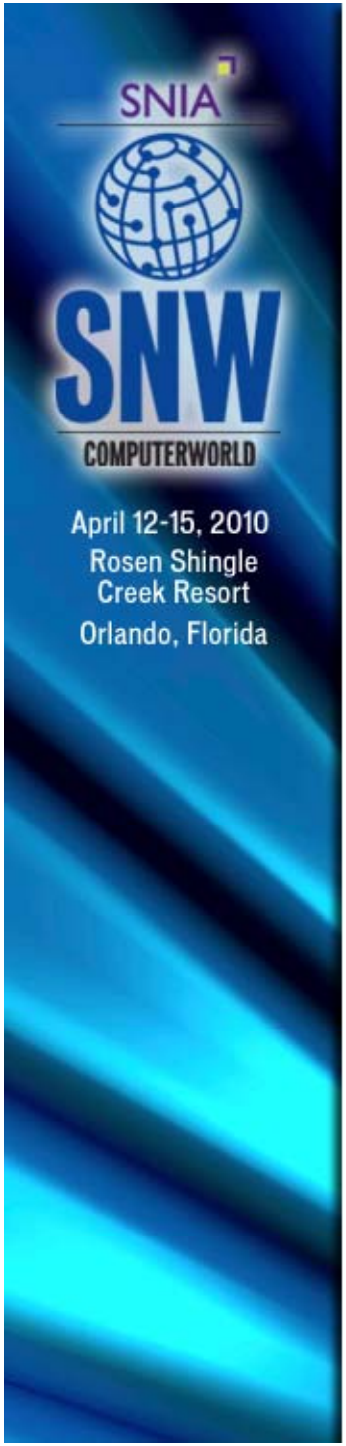
$$p(n) = (p_{\text{single-unit}})^n$$

probability of loss ($p_{\text{single-unit}} = 0.5$)





Applies to links as well as nodes



Data Versus Information

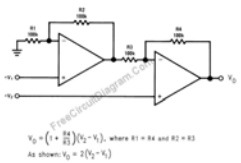
Neville Holmes
The Profession Column in IEEE *Computer* Magazine

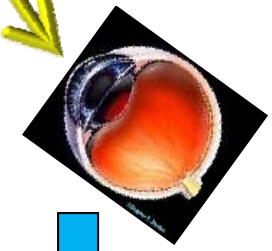
Data *represents* information

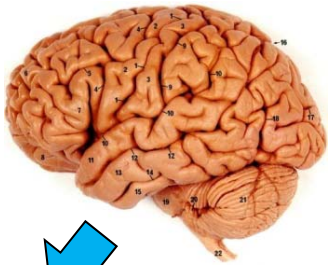
Data is the stored pattern on a medium

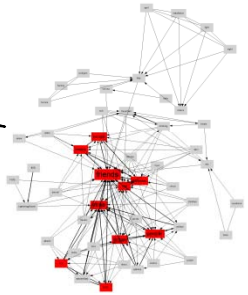
Information is how humans interpret the data and extract meaning

But it's really a continuum









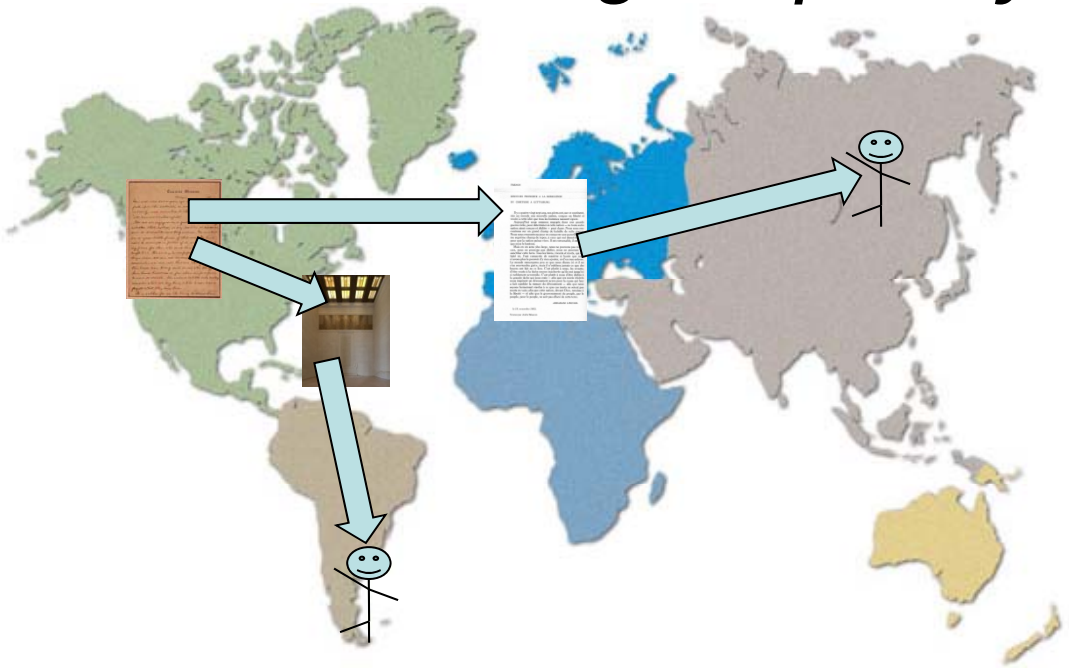
Bits on medium
Data

Series of transformations

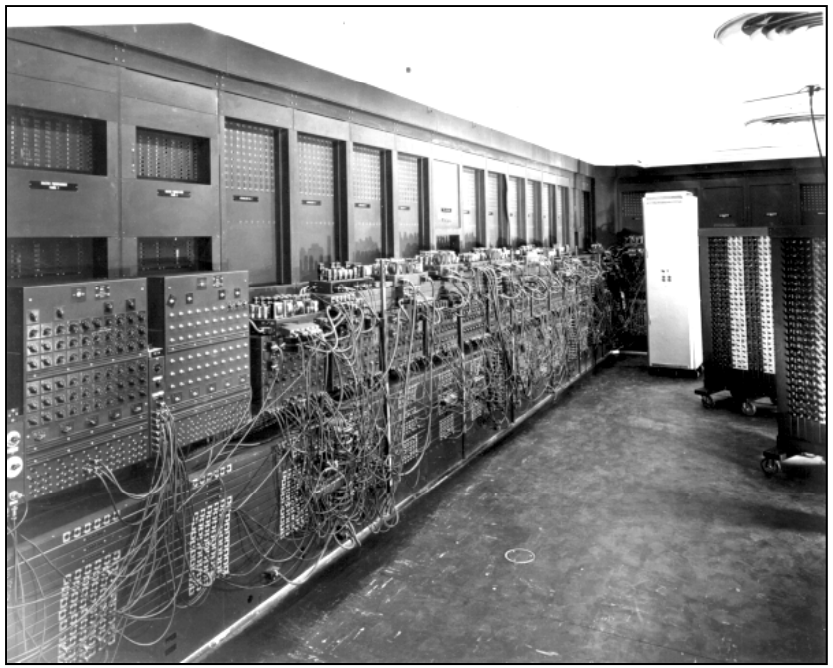
Semantic representation
Information

Principle of Continual Transformation

The quality of preservation increases as data is transformed toward information in increasing depth and with increasing frequency

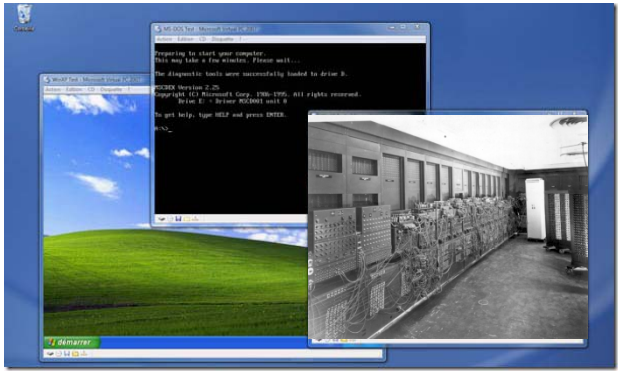


Preserving Code



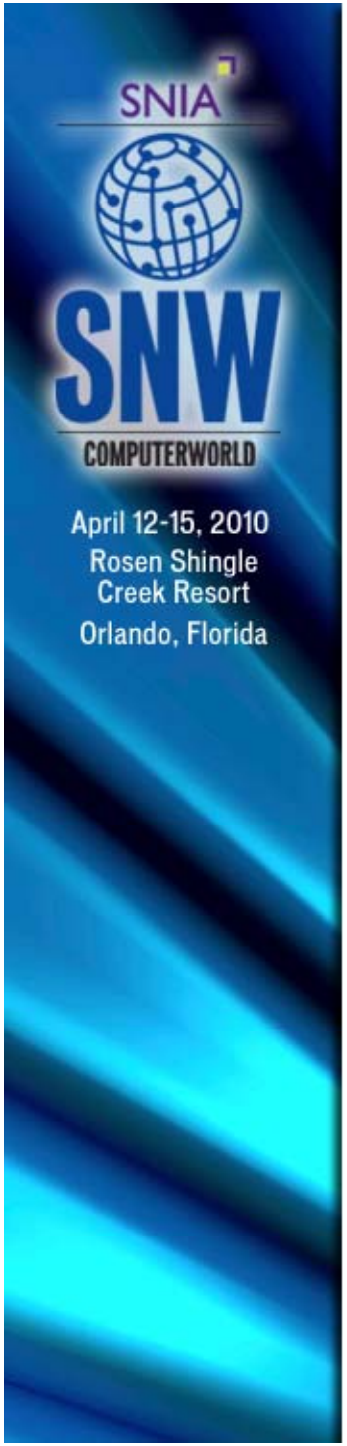
ENIAC
(circa 1946)

Escrowed Code



Virtual ENIAC

Compiled Code



All three principles must be applied to achieve good preservation

And they are interdependent as well

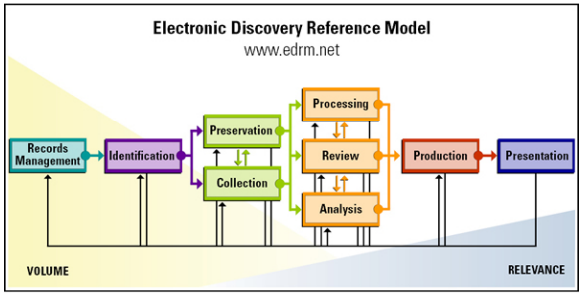
- Redundancy
  - Allows you to keep the data
- Linkage
  - Allows you to find the data
- Transformation
  - Allows you to use the data

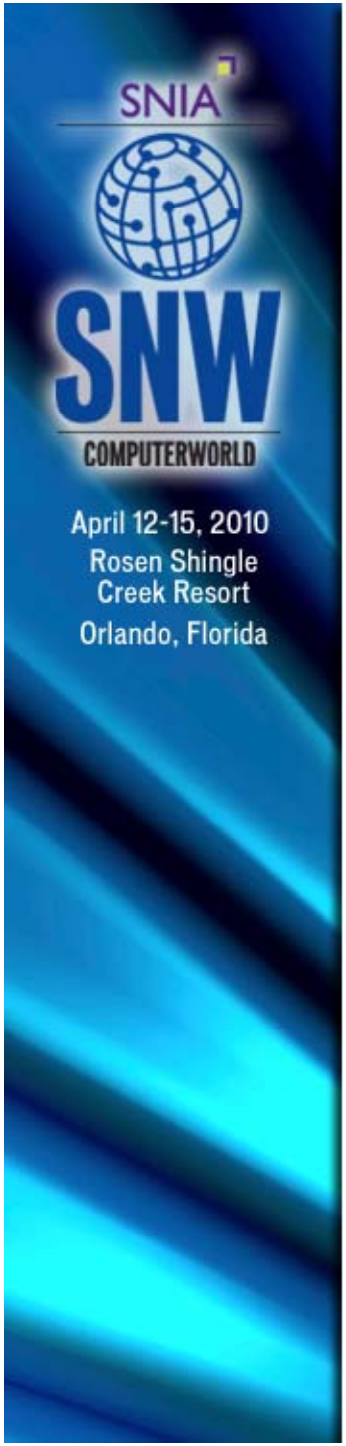
What about active data?

Compliance

Umbrella term referring to a wide range of record management practices

- Chain of custody
- Retention
- Security
- Discovery
- Some regulations, standards
  - SEC 17a-3 and 17a-4
    - WORM Email storage
  - Sarbanes-Oxley
  - EDRM (Electronic Discovery Reference Model)
- Storage points typically determined by department, person, or procedure
- Compliance applies to active and archival data



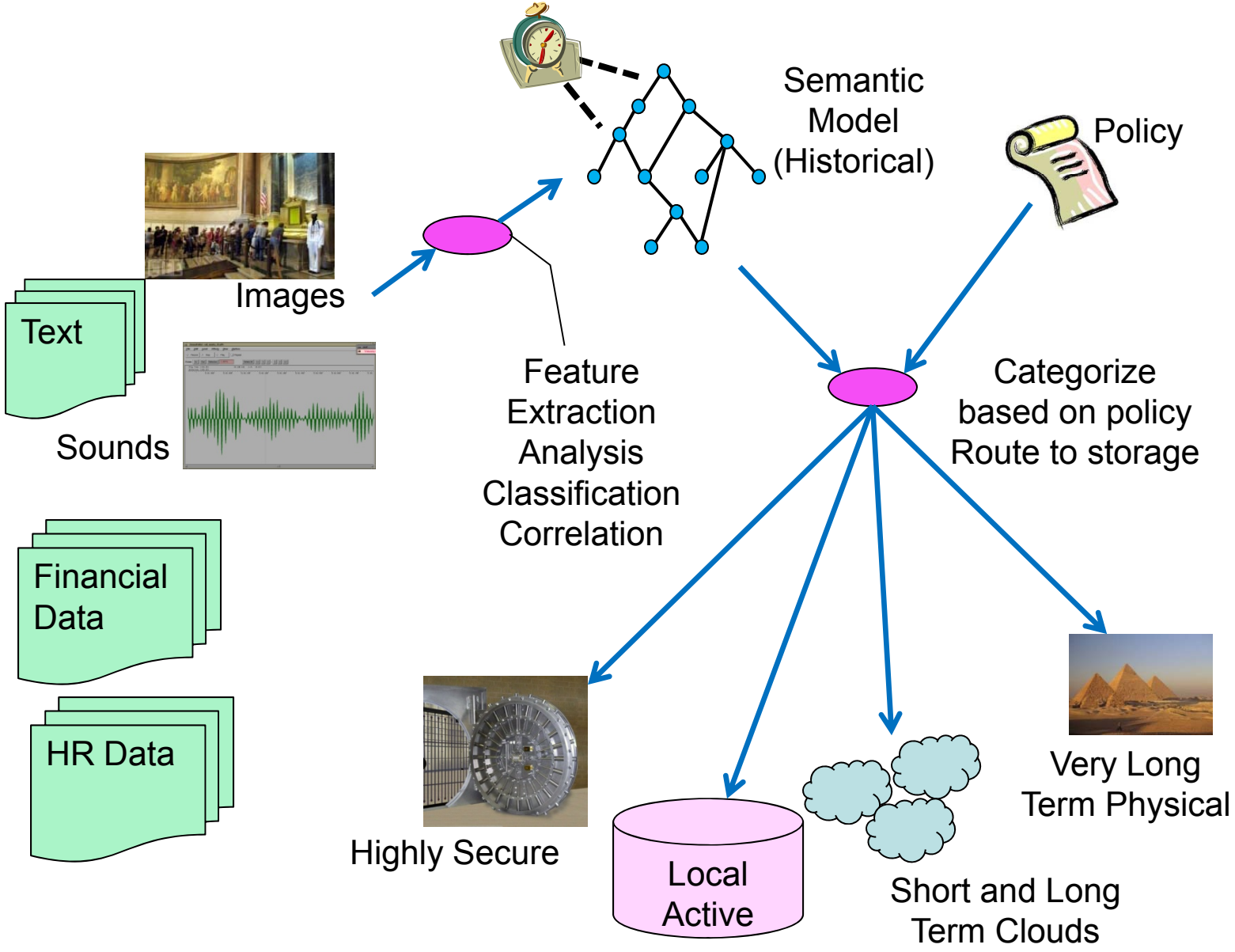


Active Data

- In day-to-day use, but…
- …same principles apply:
  - Redundancy
    - So you don't lose it
    - RAID, Backup, …
  - Linkage
    - You can lose track of it as easily as old stuff
    - Directories, search tools, …
  - Transformation
    - Operations are similar but reasons differ over the needs of preservation
    - Editing, compiling…

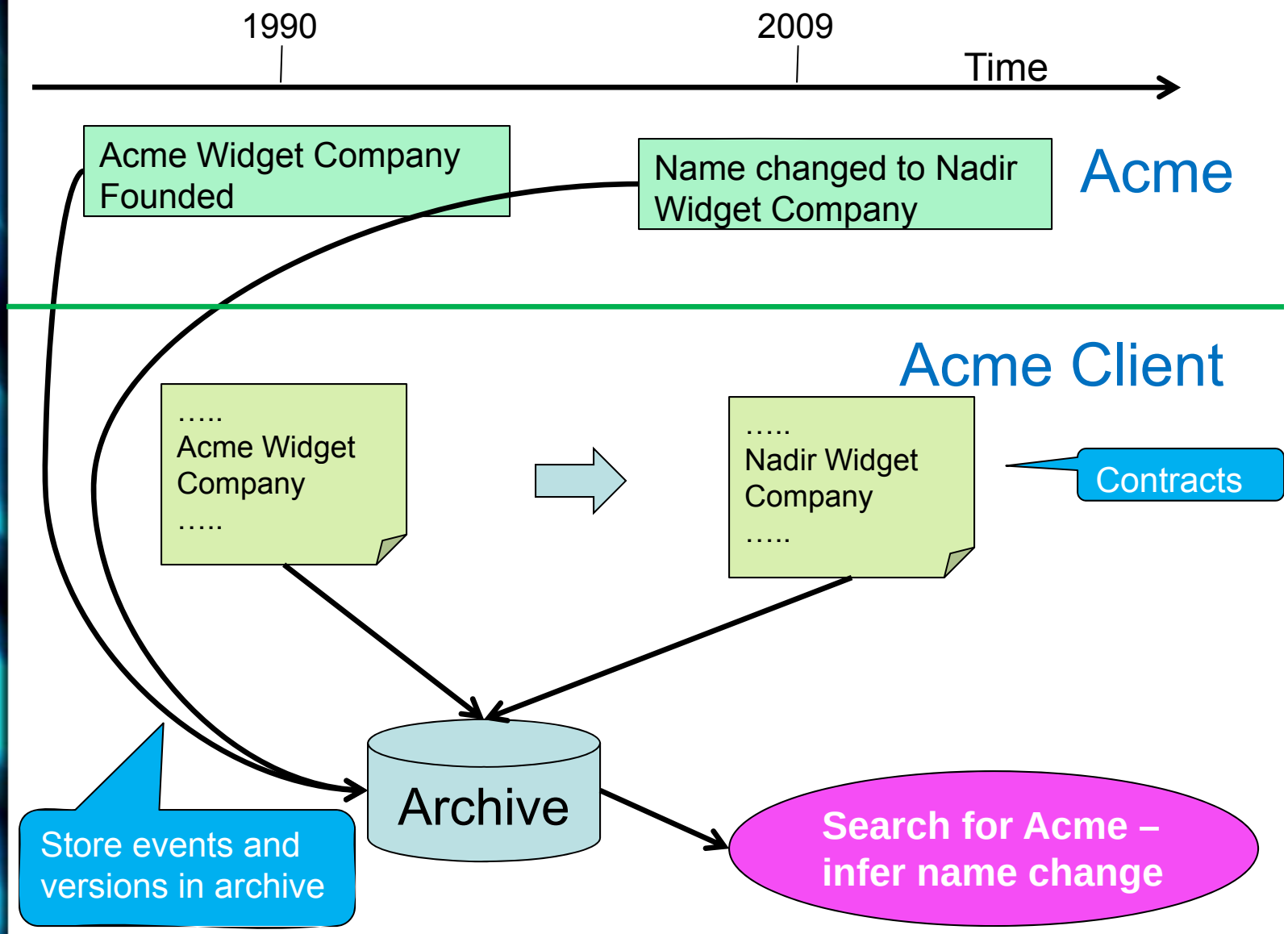
Content-driven Data Management

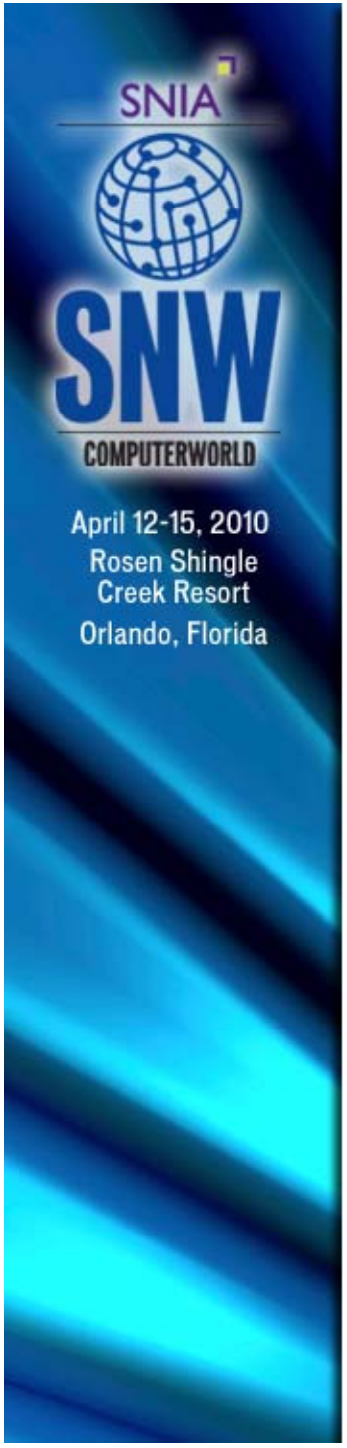
Use the information in the data to guide the management of the data



Combining Time and Text Inferences

Avoid Staleness of Data

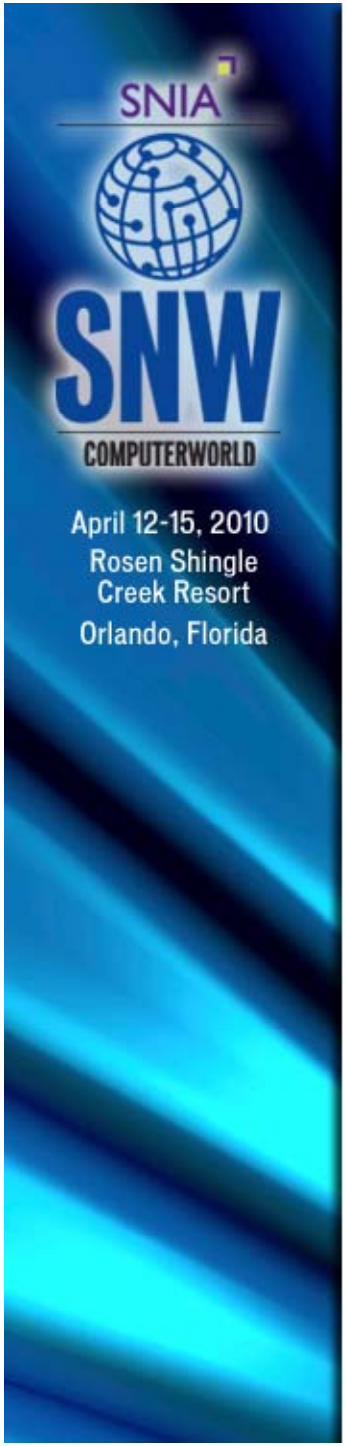




More than a Lifetime of Data and Information

- Keeping someone's life history
  - Flash keys?
    - Too easily lost
  - Implants?
    - Invasive
    - Reliability issues
    - Capacity issues
    - Obsolescence (data transformation) risk
  - Keep it in the family?
    - Eventually will get scattered, lost, destroyed

Keeping data "with the item" is not always the best way to assure the information remains available



More than a Lifetime of Data and Information

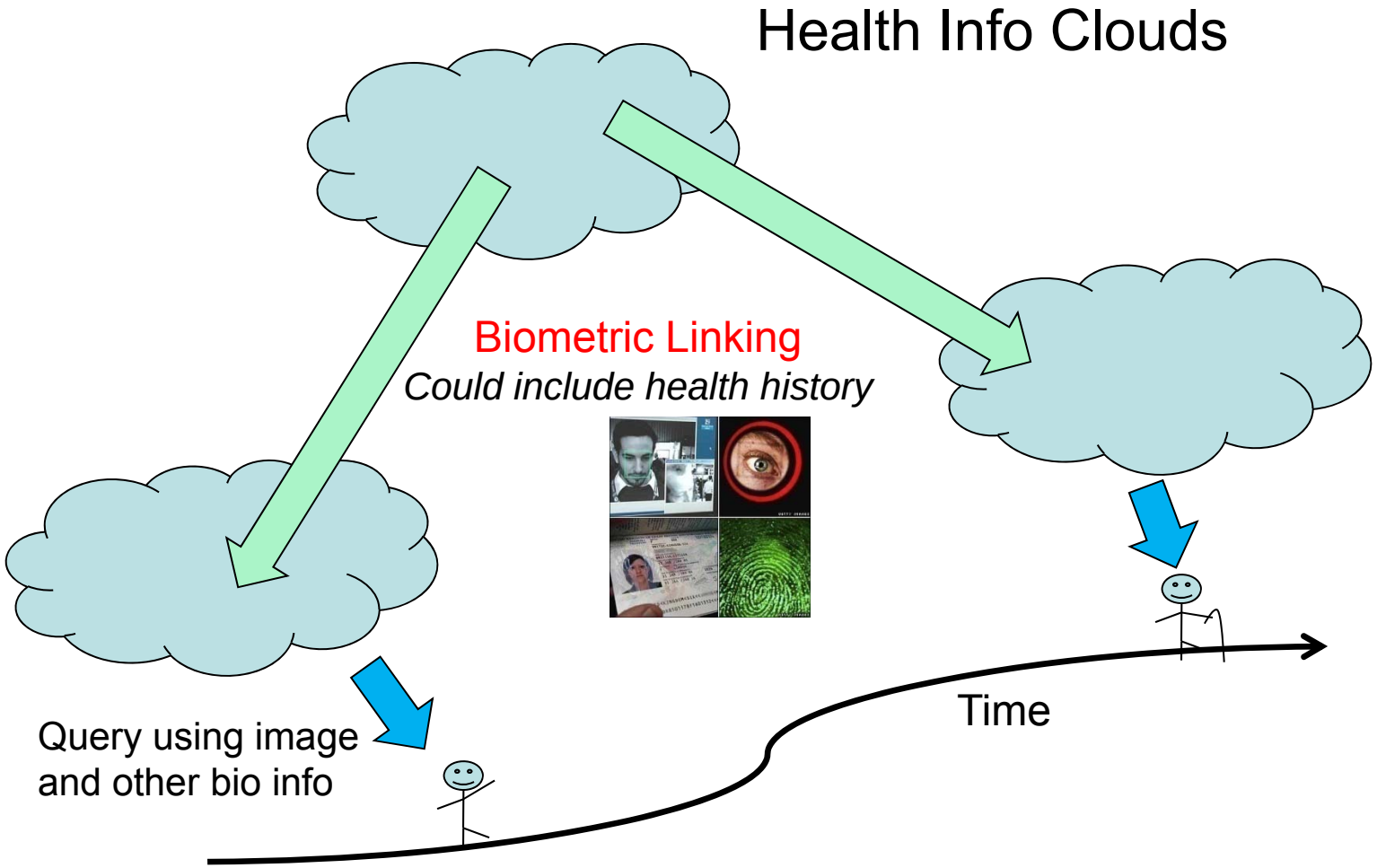
- Keep it in the cloud
  - Health, biographical, photos, videos, personal info
  - Gigabytes during lifetime
  - Megabytes beyond their lifetime
  - Compressed over time
    - Fewer people would have larger data last longer
    - Part of the transformation process
  - Some numbers:
    - ≈100 Billion people (Homo Sapiens) have ever lived on earth
    - At 1 Megabyte each 100 Petabytes
    - 1% of 100B = 10M at 1 Terabyte 10 Exabytes
    - Total current worldwide digital storage = 500 Exabytes
    - A 100 year continuous video (compressed) needs about 400 TB

It is entirely practical to imagine keeping significant history on every person indefinitely

Images and Biometrics

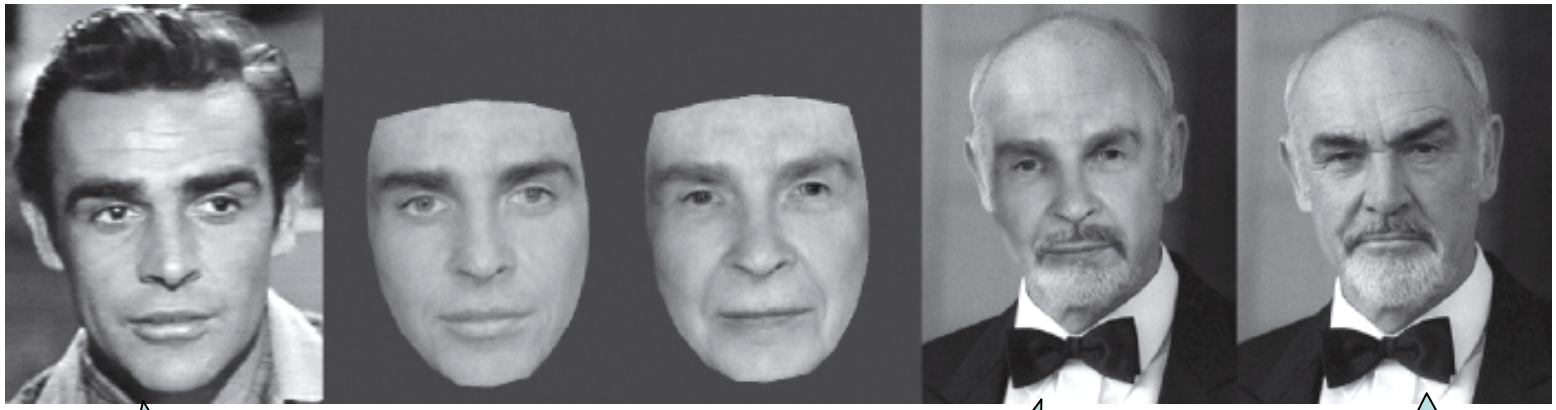
Preserve the People Too

- *Eliminate ID numbers and cards*
- *But still scary*



Predicting Sean Connery

Karl Ricanek Jr., University of North Carolina, *et al*,
***Unconstrained Biometric Identification: Emerging Technologies*,**
IEEE Computer, February, 2010

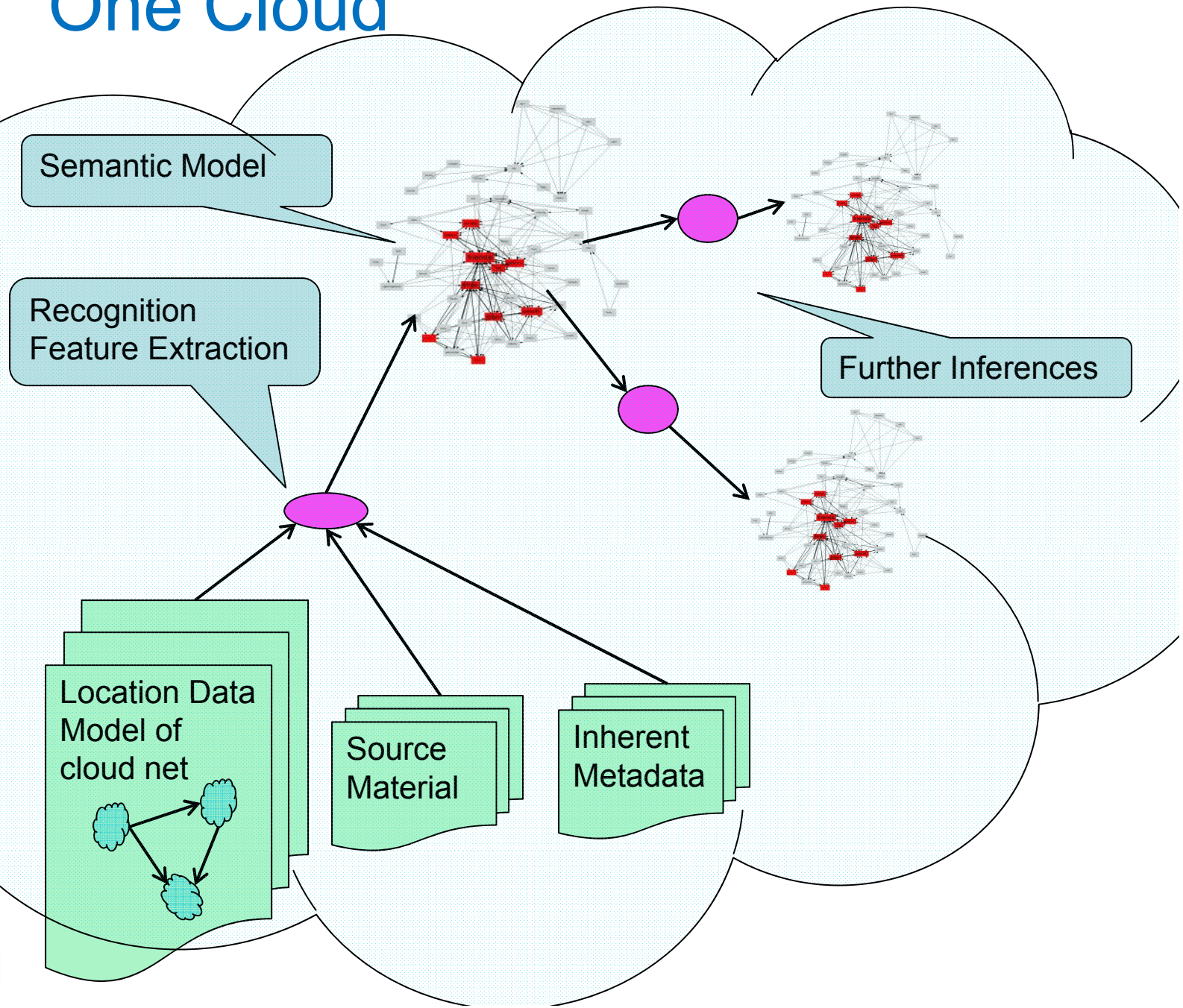


Actual photo, age 25

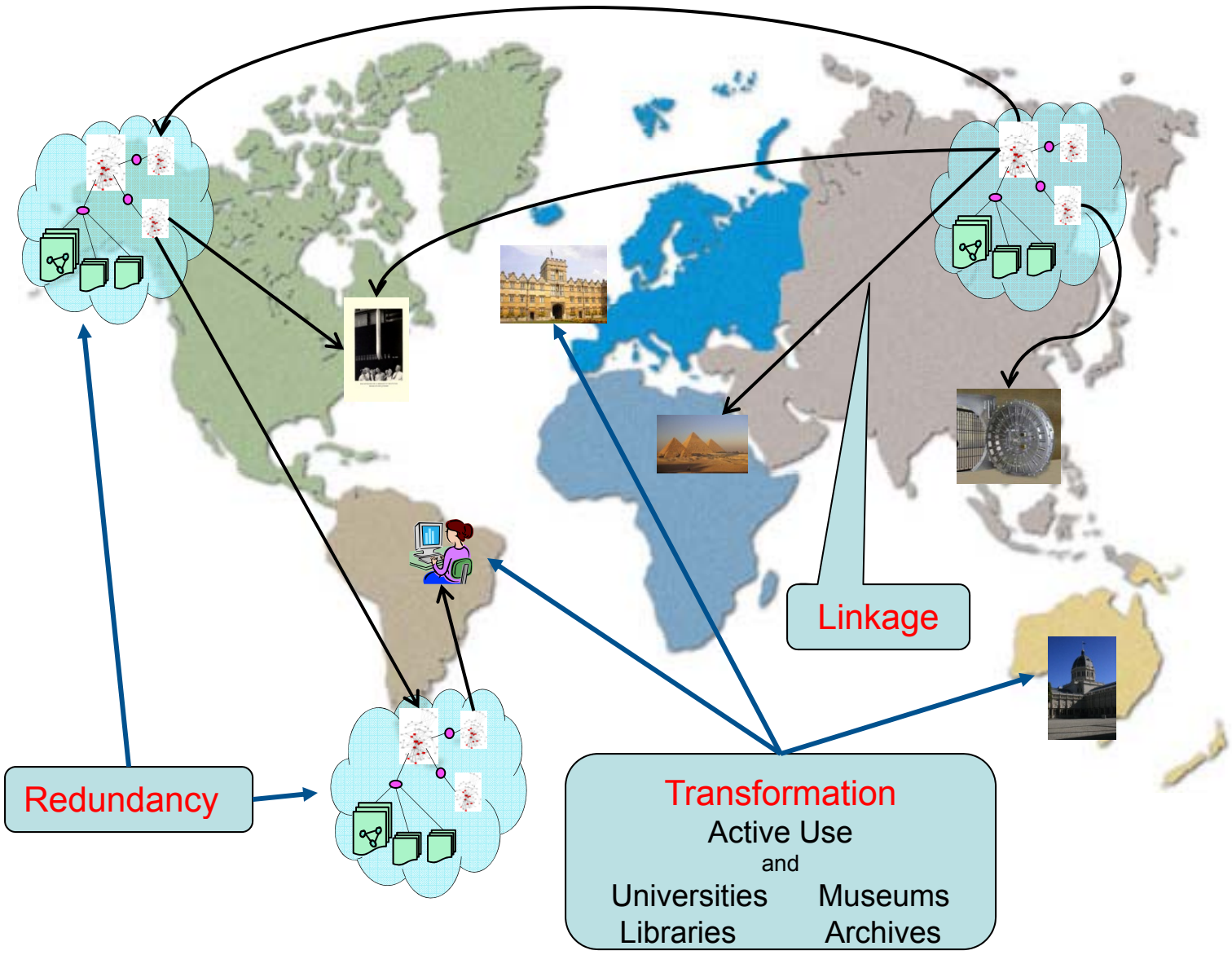
Morph algorithm prediction to age 70

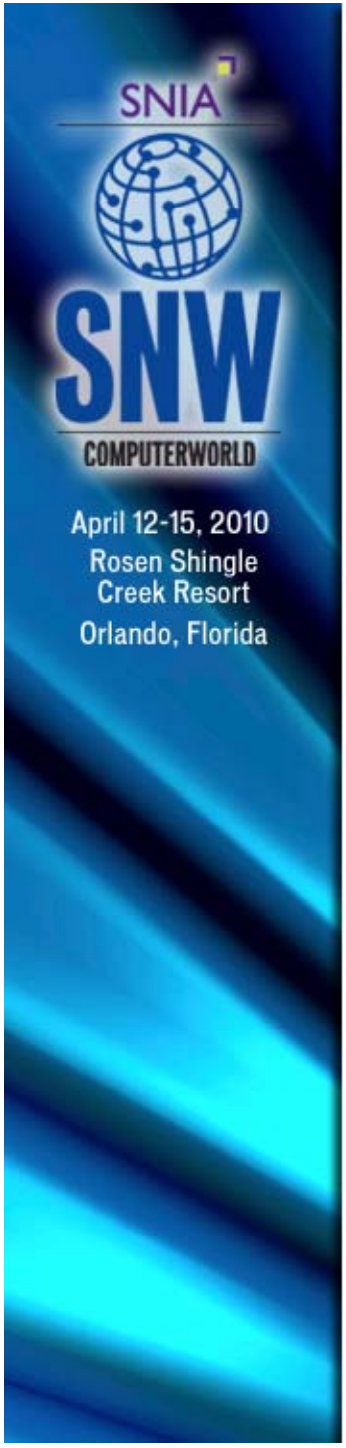
Actual photo, age 70

One Cloud



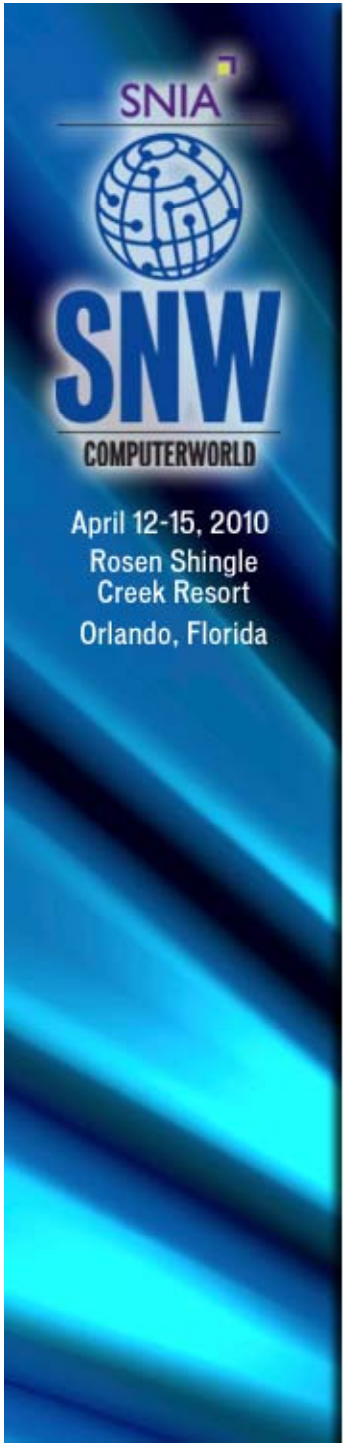
Cloud of Clouds





Challenges

- Content Models
  - Shallow, but broad, meaningful and accurate
    - Hard but easier than general AI
  - Have to account for time
  - Trust that there are few false positives/negatives
  - Some systems will always need tight management (e.g., accounting)



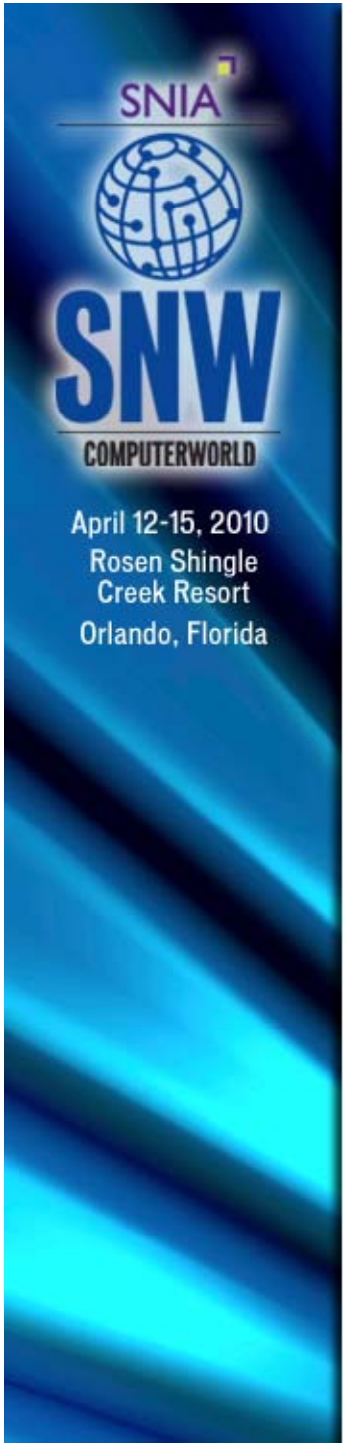
Challenges

- Interoperability
  - Basic inter-data standards *required*
    - CDMI, XAM
    - Popular vendor APIs
  - Metadata
    - A wide range of diffuse standards
    - Need a meta-meta-data standard to map them all?
      - Even simple disparities like field names and formats can cause trauma

LastWriteTime	**Local String**	**“1/15/2008 10:15 EST”**
DateTimeModified	**UTC Integer**	**123672670621**

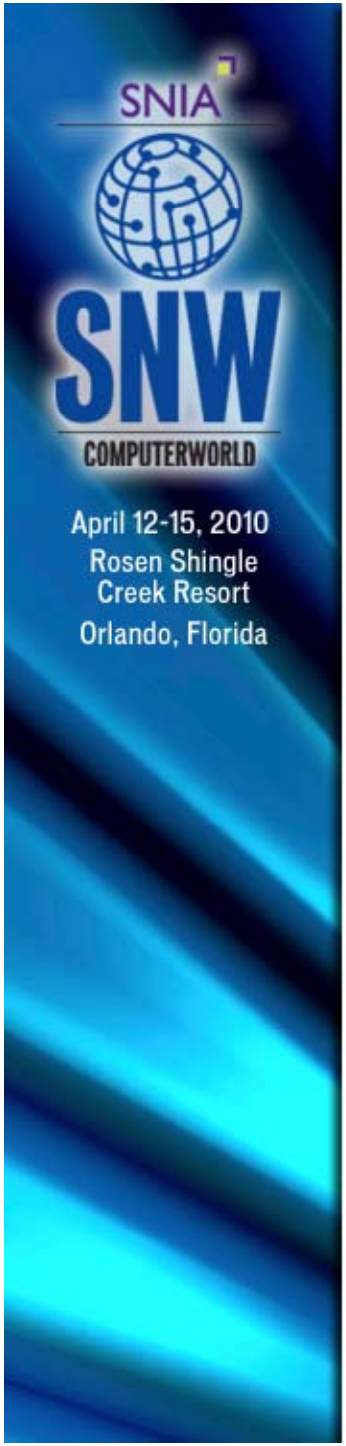
      - Can Semantic Web help?

Read these fields and understand the words



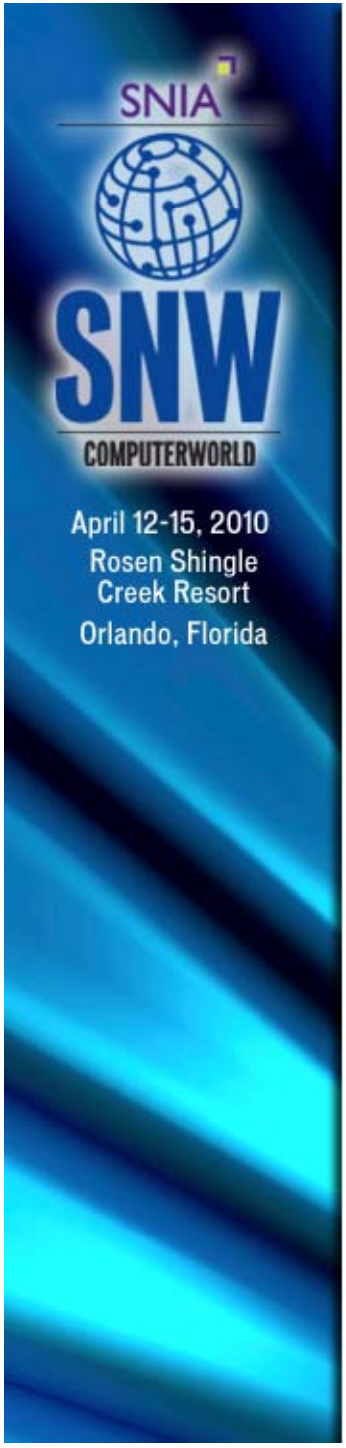
Challenges

- Transformation
  - Perhaps the most challenging of the three principles
  - Personal history over a lifetime may be easier than preserving artifacts for centuries
  - Serious efforts at automation needed here
  - Virtual Machines
    - Keep the VM up to date; underlying apps follow automatically
    - Can we apply similar ideas to physical media?
      - e.g., “universal tape drive”



Challenges

- Security, Privacy, Ethics
  - Could trump all other concerns
    - There would really be such a thing as “your permanent record”
  - Might need special crypto machines
    - Solve the homomorphism problem *practically*
    - Specialization likely – e.g., search
  - Will privacy be a concern in the future?
  - What basic ethics should be followed?
    - Who “owns” the data?
    - What rules govern its use?



Metadata Standards

A Sampling

MARC 21

MAchine Readable Cataloging
Library Science

Semantic Web

<ddi>

Data Documentation Initiative
Social Sciences

CDWA

Categories for the Description
of Works of Art

doi®

Digital Object Identifier
Invariant identification info

CSDGM

Content Standard for Digital
Geospatial Metadata

Dublin Core® Metadata Initiative
Making it easier to find information.

fgdc
Federal Geographic Data Committee

PREMIS
PRESERVATION METADATA MAINTENANCE ACTIVITY

Library of Congress

Initiatives

A Sampling

OPEN Content Alliance
Building a digital archive of global content for universal access

Non-profit private/commercial consortium for digital preservation

Semantic Web

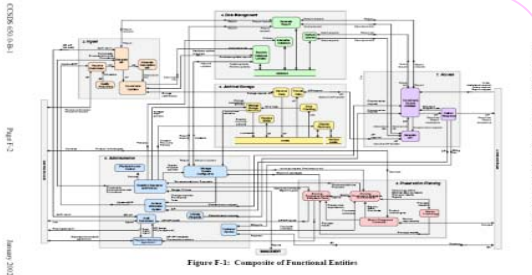
Standards for semantic representations and inference-based processing

MIT, Tim Berners-Lee

Explicit compliance-oriented records management and storage

Commercial

OAIS



ISO

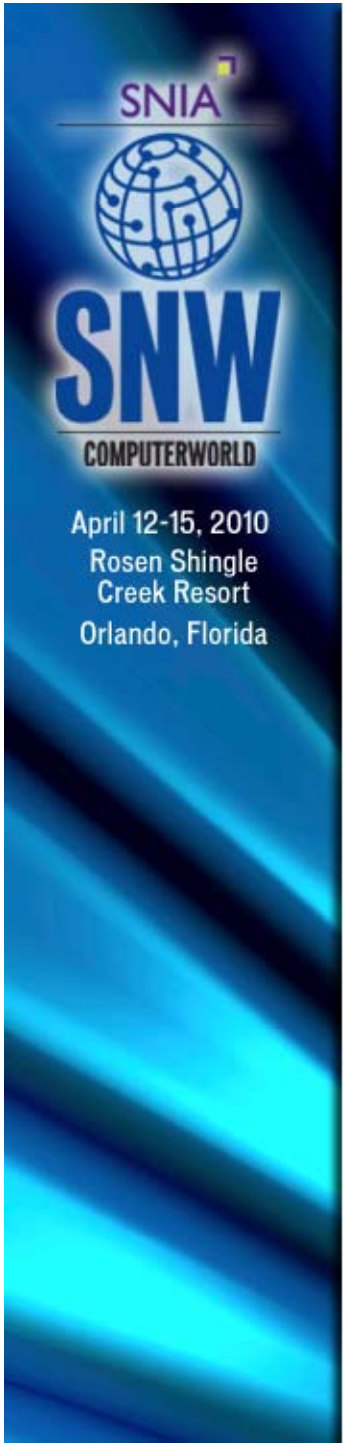
Open Archival Information System

SNIA 100 YEAR Archive Task Force

OCLC™ The world's libraries. Connected.

Online Computer Library Center

CONTENTdm digital archive software



Conclusion

- The revolution is here
- Automated interoperability is the key
- Security, privacy, ethical barriers loom
- Storage! I need more storage!