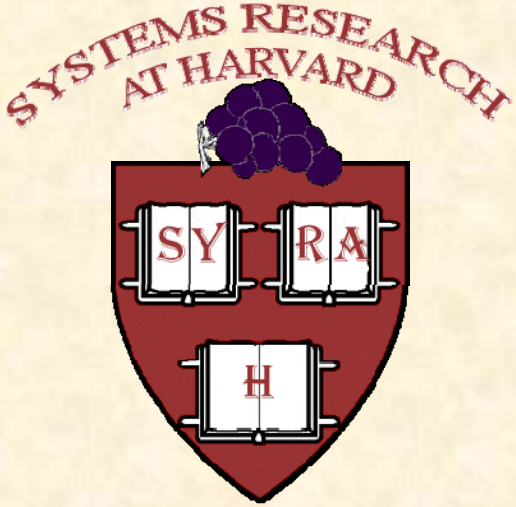
Provenance in Digital Collections



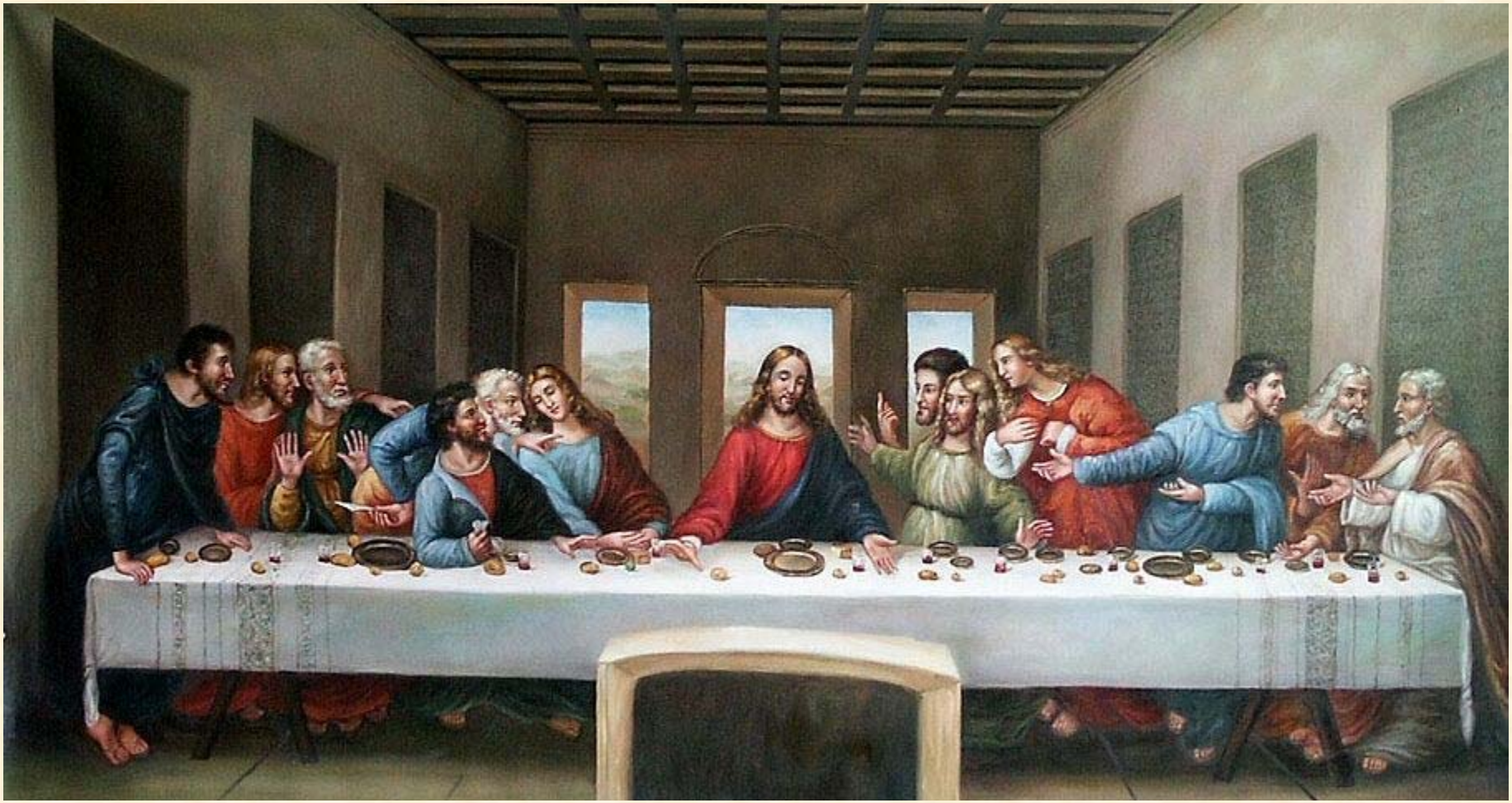
Margo Seltzer
Harvard School of Engineering and Applied Sciences

September 20, 2012

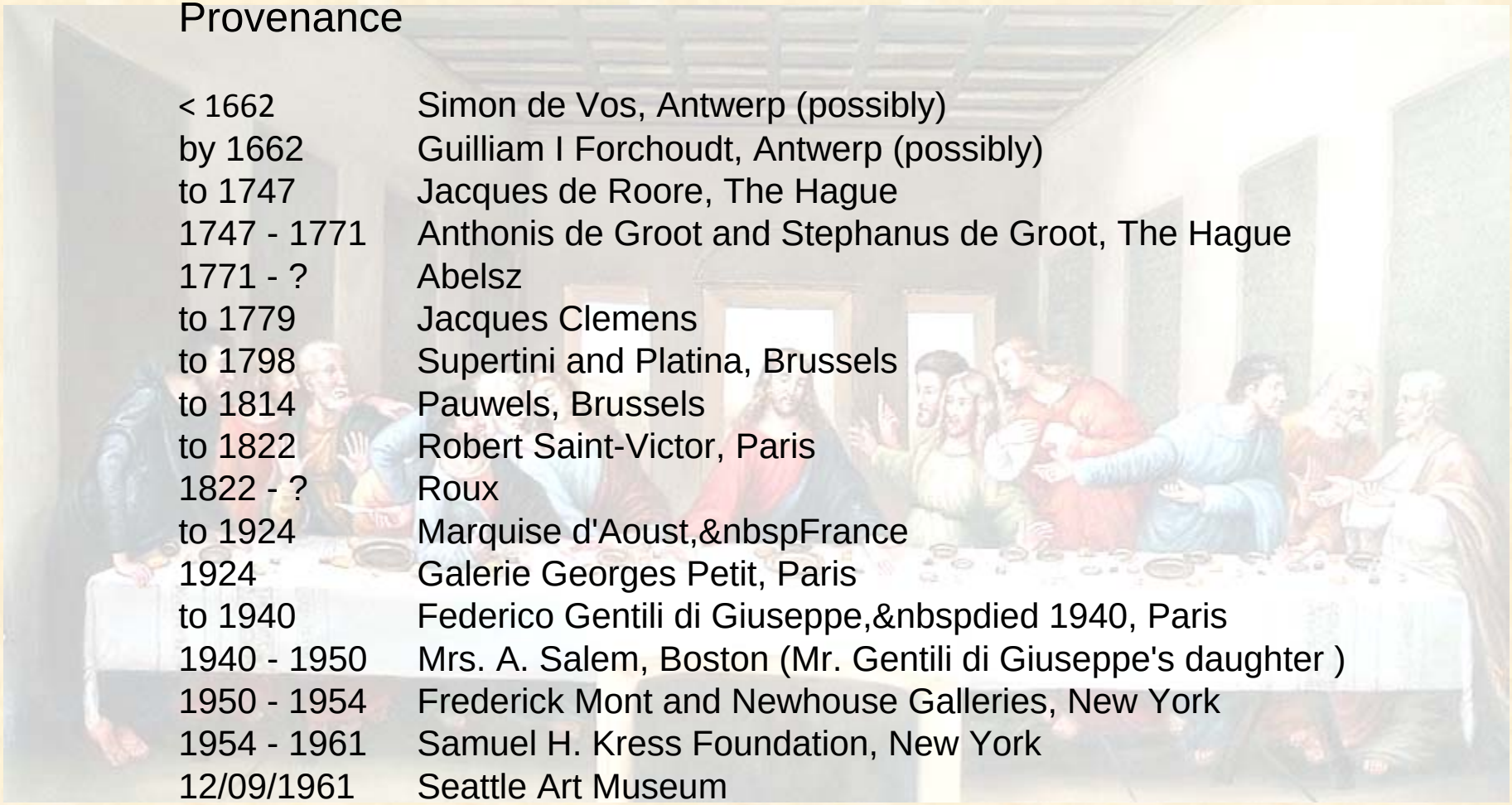
Provenance: Special Metadata

- From the French word for “source” or “origin”
- The complete history or lineage of a object
- In the art world, provenance documents the chain of ownership of an artifact.
- In the digital world, provenance records:
  - The process that created an artifact
  - The transformations applied to an artifact
  - The human and computational agents that operated upon an artifact
  - Open question: sufficient information to reproduce the artifact?

Example: Art



Example: Art with Provenance



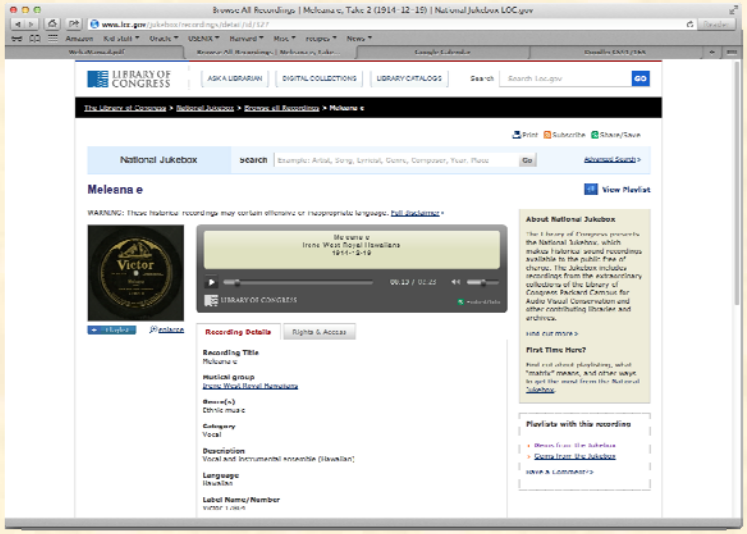
Provenance

< 1662	Simon de Vos, Antwerp (possibly)
by 1662	Guilliam I Forchoudt, Antwerp (possibly)
to 1747	Jacques de Roore, The Hague
1747 - 1771	Anthonis de Groot and Stephanus de Groot, The Hague
1771 - ?	Abelsz
to 1779	Jacques Clemens
to 1798	Supertini and Platina, Brussels
to 1814	Pauwels, Brussels
to 1822	Robert Saint-Victor, Paris
1822 - ?	Roux
to 1924	Marquise d'Aoust, France
1924	Galerie Georges Petit, Paris
to 1940	Federico Gentili di Giuseppe, died 1940, Paris
1940 - 1950	Mrs. A. Salem, Boston (Mr. Gentili di Giuseppe's daughter)
1950 - 1954	Frederick Mont and Newhouse Galleries, New York
1954 - 1961	Samuel H. Kress Foundation, New York
12/09/1961	Seattle Art Museum

Example: Data with Provenance

Meleana e

From the Library of Congress National Jukebox



From the page:

Musical Group: Irene West Royal Hawaiians
Label Name/Number: Victo 17864
Matrix Number/Take Number: B-15530/2
Recording Date: 12/19/1914
Location: Camden, NJ
Size: 10”

From http://www.loc.gov/jukebox/about/making-the-jukebox:

“A slip that provides the elements of the filename of the digital copy is inserted in the sleeve of each selected disc: institution/collection code, label name, label number, disc copy, matrix number, and take number are all noted on the slip. Using a naming convention that combines these elements allows each individual side to be fully and uniquely identified by the filename.”

More Details on the Preservation Process

- Discs are cleaned with “a mild solution of Tergitol and distilled water”
- Barcode created and attached
- “Since neither speed nor groove size were precisely consistent in the acoustical era, the initial stylus and speed setting in the studio is a best guess estimation. Because the discs coming to the studio are typically batched in label number series, they may also have similar playback speeds and groove sizes. So the engineers usually start with the settings that were used in the previous transfer then use their trained ears and studio experience to determine what changes are needed.”

What would you do if you wanted to accept contributions from others?

Where Does Provenance Come From?

- From instruments: barcode readers, audio devices, thermometers, cameras, telescopes, sensors, …
- From software: Photoshop, your database, your home-grown tools, the network
- From system software: the operating system, libraries, kernel modules
- From tools: the compiler, the interpreter, your source code control system.
- In other words: from lots of places and is the result of both automatic and manual data manipulation.

The Vision for a Digital Collection

- **All** data has provenance.
- **Applications** generate provenance.
- **Systems** generate provenance.
- **Users** generate provenance.
- Provenance:
  - Is tamper proof
  - Can be authenticated
  - Can be queried and searched
  - From different systems can be treated uniformly

Why is this hard?

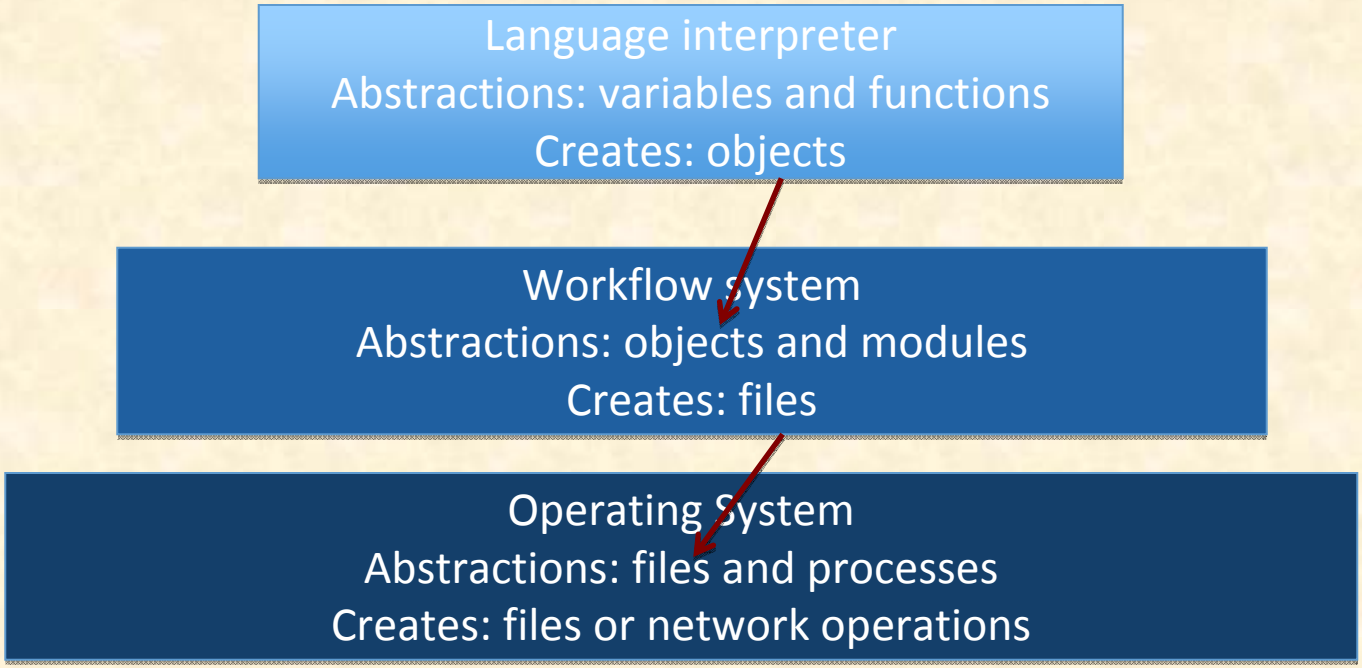
- Most systems don't record provenance.
- Even the ones that do have a myopic perspective; they refer to their own abstractions and do not provide a way to connect their abstractions with those of other layers of software.
  - Operating system: files
  - Database systems: tuples
  - Workflow engines: objects
  - Applications:
    - Tangible artifacts (e.g., a particular disc)
    - Ephermeral objects (e.g., inks or sessions from a browser)
    - Pieces of data (e.g., a paragraph from a word processor)
- Interoperability is lacking
  - Each system knows about its native objects.
  - Lacks understanding of what happens in black boxes.
  - Lacks connections with things that happen outside of It.

Doesn't the new W3C Provenance standard give us what we need?

- Good starting point:
  - Standard terminology and data model
  - "Alternate" mechanism provides a way to bridge between different collection agents
- Limitations
  - Web-focused (with goal to be broadly applicable)
  - Assumes that systems translate to a lingua franca for exchange:
    - "A pragmatic approach is to consider a core provenance language with an extension mechanisms that allow any provenance model to be *translated* into such a lingua franca and *exchanged* between systems."
    - This loses semantic meaning
  - Assumes immutable objects and does not model "versions"
    - Hides challenges in modeling changes to collections.
  - Ignores issues of authenticity and security
  - Assumes RDF-style representation and SPARQL for query.

Explicit support for Layering & Integration

- Key concept:
  - Each layer collects provenance.
  - Each layer associates its objects with objects in its adjacent layers.



Making Layering Work

- Can't we just place all provenance in a central repository?
  - All participants would need to agree on naming conventions.
  - Participants would need to be able to generate references to objects created by other participants.
  - What happens when you add a new participant with a new naming mechanism?
- In layering, a participant discloses the relationship between its objects and those in the layer below; that layer then becomes responsible for further transmission.

Layering provides a natural way to
transmit and integrate provenance
and facilitates query across the layers.

Examples of Layered Systems

- We've built a provenance-aware storage system (PASS).
  - Layers on NFS and/or a cloud storage service.
  - Enables Kepler workflow engine to layer on top of it.
- We prototyped simple database provenance in SQLite
  - Layered on top of PASS
  - (Did the third provenance challenge with it.)
- We have a provenance-aware python workflow engine (Starflow).
  - Layers on PASS
  - Provides auto-update capabilities
  - Integrates with StarCluster
- Other possibilities:
  - Provenance-aware R
  - Provenance-aware browsing

From here to there

- Build provenance into systems from day one.
- Or add provenance to legacy systems.

JUST DO IT

- With layering, standards can be simpler.
- Each layer focuses on creating and transmitting provenance it understands.
  - Associate entities in one layer with entities in another layer (explicitly).
  - Queries expressed in terms of whatever layer and name make sense for the querier.

Provenance: Just Do It

Margo Seltzer

margo@eecs.harvard.edu

http://www.eecs.harvard.edu/~syrah/pass