Involuntary Secondary Permanence:
Do many copies replace the one original?

Luciana Duranti
Yale University Library
4 November 2014

Archival Concepts

Archival concepts are grounded in Roman Law

- Documents preserve perpetual memory of the facts and acts to which they relate
- Authentication is based on respect of procedure in documents' creation and use
- Deposit of documents in a public place guarantees their reliability as testimony
- Antiquity provides documents with the highest authority (as they could not have been generated to serve the interests for which they are used today)
- Unbroken legitimate custody ensures documents' authenticity (by inference)

(Justinian Code A.D. 565)

Diplomatics

The rule of law is easily circumvented: the trustworthiness of records needs to be tested using scientific methods.

Diplomatics (1681) , Dom Jean Mabillon

Trustworthiness of documents has to be based on the process of their genesis, on their characteristics of form and structure, and on their transmission through time and space.

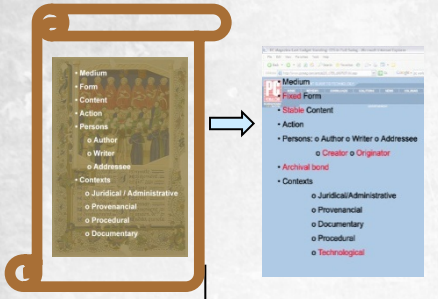
The **Bella Diplomatica** (judicial disputes on authenticity of documents) based on diplomatic methods gave origin to the **Law of Evidence**

By mid 18th century all faculties of law in Europe taught archival science and diplomatics as "forensic" disciplines

Archival Diplomatics

Dr. Luciana Duranti
The University of British Columbia

The Concept of Record



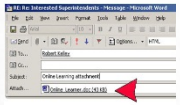
Digital Record Characteristics

On the face Of the Record

Formal Elements

Attributes

Digital Components



Lifecycle of Digital Records
Phase 1: Records of the creator
Phase 2: Authentic copies of the records of the creator

Luciana Duranti
Email: luciana.duranti@ubc.ca
www.ciscra.org

Archival Diplomatics

The integration of archival and diplomatic theory about the genesis, inner constitution, and transmission of documents; and about their relationship with the facts represented in them, and with other documents produced in the course of the same function and activities, and with their creators.

Retrospective Use

Prospective Use

Dynamic and Interactive Records
Stable Content
Fixed Documentary Form
Bounded Variability

Genesis of the Digital Records Workflow: actio et conscriptio

Initiative

Inquiry

Consultation

Deliberation

Deliberation Control

Execution

Functions of Records
Probative/Dispositive
Supporting/Narrative
Instructive/Enabling

Categories of Records

- Manifested:
- Stored:

} Form, Content, and Composition Data

Application: Research Projects
UBC Project (1994 - 1997)
InterPARES 1 (1999 - 2001)
InterPARES 2 (2002 – 2006)
InterPARES 3 (in application)

Status of Transmission
Draft
Original
Authenticated original
Copy (e.g., authentic copy)

The Concept of Trustworthiness

Reliability

The trustworthiness of a record as a statement of fact. It exists when a record can stand for the fact it is about.

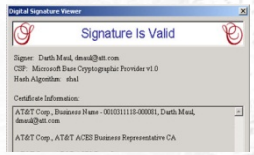
Authenticity

- identity
- integrity

The trustworthiness of a record as a record; i.e., the quality of a record that is what it purports to be and that is free from tampering or corruption.

Accuracy

The degree to which data, information, documents or records are precise, correct, truthful, free of error or distortion, or pertinent to the matter.



Metadata
Identity Metadata
Integrity Metadata

✓As a Means of Authentication

Authentication:
A means of declaring the authenticity of a record at one particular moment in time

InterPARES Trust

Trustworthiness

Reliability

The trustworthiness of a document as a statement of fact,

based on:

- the competence of its author
- the controls on its creation

Accuracy

The correctness and precision of a document's content

based on:

- the competence of its author
- the controls on content recording and transmission

Authenticity

The trustworthiness of a document that is what it purports to be, untampered with and uncorrupted

based on:

- identity
- integrity

InterPARES Trust

Status of Transmission

Degree of perfection of a document:

- **Draft** – a document prepared for purposes of correction and meant to be provisional, temporary
- **Original** – the first, complete document capable of reaching the purposes for which it was intended
  - Primitiveness, completeness, effectiveness (e.g. fax)
- **Copy** – a reproduction of another document, which may be an original, a draft or another copy
  - Authentic copy, facsimile (can be made by anyone), copy in the form of original, imitative, simple, insert, inspeximus (vidimus)

InterPARES Trust

...in the Digital Environment

- **Digitized originals** on other media correspond to a traditional authentic copy, are stable, and every instantiation that is authenticated has the force of original
- **Born digital originals** exist for a nano-second
- Synchronic and diachronic copies are generated for redundancy or distribution but they are not identical copies
- Preserving a born digital document involves maintaining the ability to re-produce it or re-create it
- But a document, as well as its copies, to be such, must have fixed form and stable content

InterPARES Trust

Document with Fixed Form

- A digital entity has fixed form if its binary content is stored so that the message it conveys can be rendered with the **same documentary presentation** it had on the screen when first saved (even if different digital presentation: Word to PDF)
- An entity has fixed form also if the same content can be presented on the screen in several different ways in **a limited series of possibilities:** we have a different documentary presentation of the same stored entity having stable content and fixed form (e.g. statistical data viewed as a pie chart, a bar chart, or a table)

InterPARES Trust

Document with Stable Content

- An entity has stable content if the data and the message it conveys are **unchanged and unchangeable**, meaning that data cannot be overwritten, altered, deleted or added to

- **Bounded Variability**: when changes to the documentary presentation of a determined stable content are limited and controlled by fixed rules, so that the same query or interaction always generates the same result, and we have different views of different subsets of content, due to the intention of the author or to different operating systems or applications

Digital Document Parts

- **Formal Elements**: constituent parts of the documentary form as shown on the document's face, e.g. address, salutation, preamble, complimentary close
- **Metadata**: the attributes of the document that demonstrate its identity and integrity
- **Digital Components**: stored digital entities that either contain one or more document or are contained in the document and require a specific preservation measure

Stored and Manifested Document

- **Stored document**: it is constituted of the digital component(s) used in re-producing the document, which comprise the data to be processed in order to manifest the document (content data and form data) and the rules for processing the data, including those enabling variations (composition data)

- **Manifested document:** it is the visualization or instantiation of the document in a form suitable for presentation to a person or a system. Sometimes, it does not have a corresponding stored document, but it is re-created from fixed content data when a user's action associates them with specific form data and composition data (e.g. a document produced from a relational database)

InterPARES Trust

Types of Digital Documents

Static: They do not provide possibilities for changing their manifest content or form beyond opening, closing and navigating: e-mail, reports, sound recordings, motion video, snapshots of web pages

Interactive: They present variable content, form, or both, but the rules governing the content and form of presentation are fixed

Digital Copies

- In light of the above, it is clear that, in the absence of originals, we need to either make or identify the most **authoritative** copy as the master copy for preservation
- Authenticity Metadata are the primary means of providing or identifying the authority of a copy

InterPARES Trust

Authenticity: Identity

The whole of the attributes of a document that characterize it as unique, and that distinguish it from other documents.

Identity metadata:

•names of the persons concurring in its creation

•date(s) and time(s) of its genesis, issuing and transmission

•the matter or subject, or the action in which it participates

•the expression of its relationships to other documents

•documentary form name

•digital presentation (format)

•the indication of any attachment(s)

•digital signature (if applicable)

Authenticity: Integrity

A document has integrity if the message it is meant to communicate in order to achieve its purpose is unaltered.

Integrity metadata:

- name(s) of handling persons over time
- name of person responsible for keeping the record
- indication of annotations
- indication of technical changes
- indication of presence or removal of a digital signature
- time of planned removal from the system
- time of transfer to a custodian
- time of planned deletion
- existence and location of copies in or outside the system

InterPARES Trust

Integrity

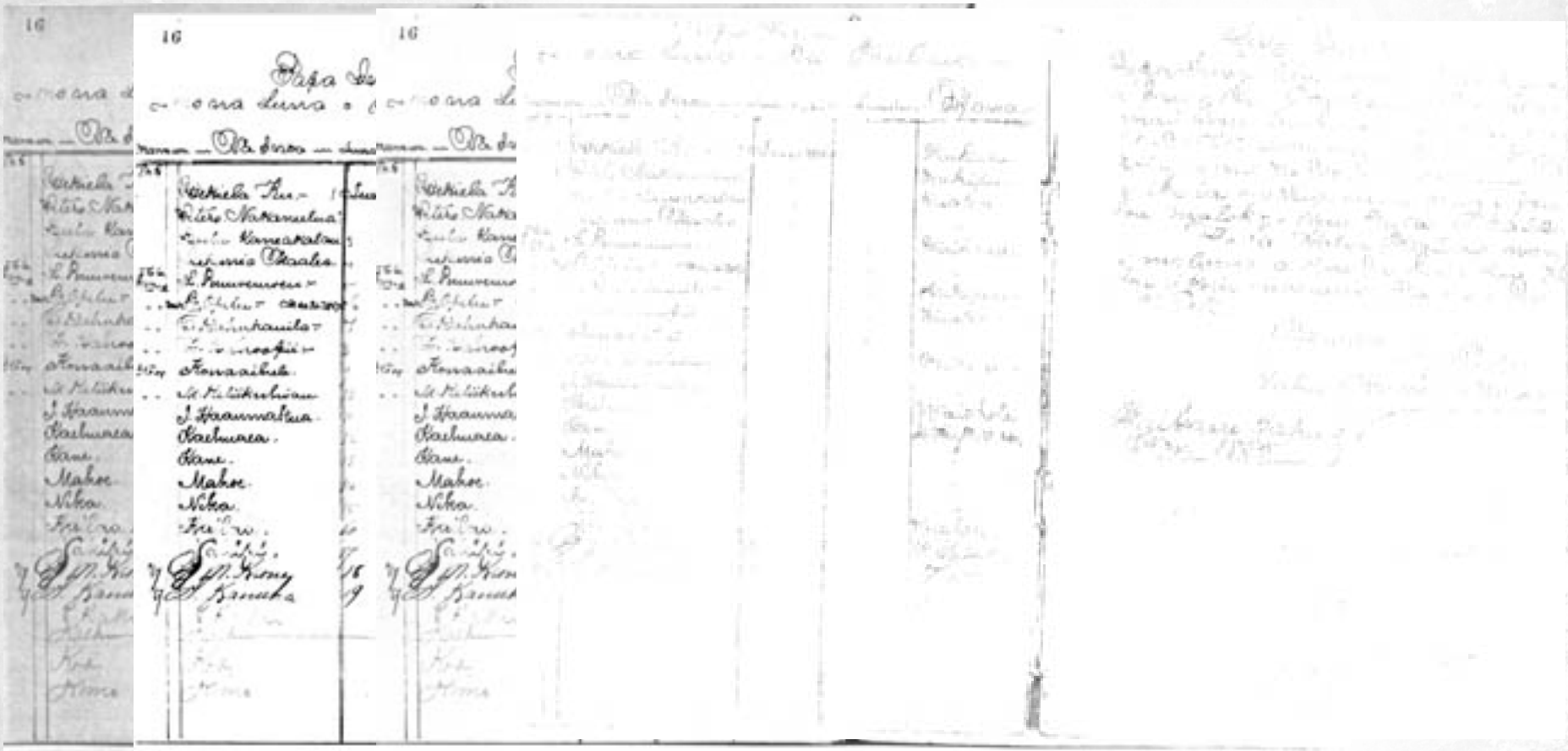
The quality of being complete and unaltered in all **essential** respects.

We were never fussy about it. What if a document had holes, was burned on a side or the ink passed through?

The same definition of integrity was used with respect to data, documents, records, copies, records systems

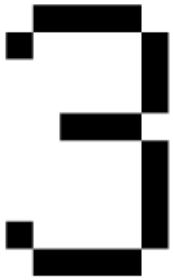
As long as it was good enough to understand it, ...but how good is good enough in the digital environment?

Loss of Integrity: Analog Document



Loss of Integrity: Digital Document

- If Original Bits 101
- Change state to 110
- Continues to a 011



- Same bits, but

  Different value

Data Integrity

Based on **Bitwise Integrity**:

- the fact that data in the document are not modified either intentionally or accidentally "without proper authorization."
- The original bits are in a complete and unaltered state from the time of capture, that is, they have the exact and same order and value

A small change in a bit means a very different value presented on the screen or action taken in a program or database.

Protecting Documents From Data Alteration

- **Intentional alteration** is preventable through permission and access controls, and strong methods like Checksum and HASH Algorithms
- **Accidental alteration** is preventable through additional hardware and/or software
- We also need methods of determining whether the document has been altered, maliciously or otherwise, and cannot rely on file size, dates or other file properties
- **We need logs**: sets of files *automatically* created to track the actions taken, services run, or files accessed or modified, at what time, by whom and from where

InterPARES Trust

Duplication Integrity

Duplication integrity: it means that, given a document, the process of creating a copy does not modify its data (either intentionally or accidentally), and the output is an exact bit copy of the original data set (form, content and composition data).

Duplication integrity is also linked to time and one should consider the use of time stamps for that purpose.

But, in the digital environment, when we say duplication, we need to be explicit about what we mean...

Documentary Duplication: Copy

Copy: selective duplicate (e.g. PDF)

- You can only copy what you can see
- Rarely includes confirmation of completeness
- Provides incomplete picture of the digital environment

Forensic Duplication: Image

Image: a bit by bit reproduction of the storage medium and its content, including ambient data, swap space and slack space

A full copy of the data on a storage device can be done regardless of operating system or storage technology

Authentic Duplication

Increasingly, in legal proceedings, **native formats** are to be submitted as authentic copies, keeping in mind the definition of it.

- PDF is a file format for delivering page-based documents in a platform independent manner, preserving the appearance of the document when viewed across multiple architectures.
- If the accurate display of fonts is an evidentiary requirement, PDF/A is preferable. Although it does not allow audio/video content, JavaScript, compression, and encryption, it requires that all fonts be embedded and uses Adobe's Extensible Metadata Platform (XMP) metadata rules with the ability to supply new metadata schema if needed.

InterPARES Trust

Authentic Duplication (cont.)

- PDF/A-2 increases accessibility, includes improvements for smaller file sizes, permits the use of JPEG2000 image compression, and allows the attachment of other PDF/A files.

- PDF/A-3 has the same functionalities as PDF/A-2 but, in addition to other PDF/A files, can embed any kind of data stream. Document viewers designed to work with the specification will display the record content just as with PDF/A-2, but can have an additional recommended functionality where, at the user's request, the embedded data can be extracted from the PDF and used/opened in any desired manner.

Authentic Duplication for Conversion and Migration

It appears that the PDF/A-3 format would support evidentiary requirements from both an evidentiary and an authentic preservation perspective when

- addressing issues of obsolescence or
- improving readability/usability across platforms

because

- the static visual elements of the main display document present the content with stability and the form with fixity, and
- a larger contextual metadata set (or sets) could be stored in the 'dumb' data sections to ensure verification of authenticity, and
- any concern about "best evidence" or data integrity can be addressed with the embedding of the original bit-stream of the source itself.

InterPARES Trust

Process Integrity: Principles

The authority of a copy depends also on process integrity

Principle of Non-interference: the method used to re-produce or re-create a digital document does not change the digital entities

Principle of Identifiable interference: if the method used does alter the entities, the changes are identifiable and identified

These principles embody the ethical and professional stance of a neutral third party, the trusted designated custodian

Authentication

A means of declaring the authenticity of a document at one particular moment in time

Example: the **digital signature**. Functionally equivalent to medieval seals (not signatures):

- verifies origin of the document (identity)
- certifies intactness of the document (integrity)
- makes the authorship or ownership indisputable and incontestable (non-repudiation)

The analogy is not perfect, because the medieval seal was associated exclusively with a person, while the digital signature is **associated with a given person and a specific document**, and because the former is an expression of authority, while the latter is only a mathematical expression

InterPARES Trust

Digital Signature

- Cryptographic digital signatures are said to provide incontrovertible mechanisms for verifying the authenticity of digital objects
- They have been given legal value by legislative (e.g., EU: European Directive on electronic signatures) or regulatory bodies (SEC: hash algorithmes).
- Digital signatures are enabled through complex and costly public-key infrastructures (PKI) and are based on the same mathematical techniques as encryption
- **A digitally signed copy only proves that the accessed document is the same that was signed and transmitted by the person linked to the signature, not that it is an authentic copy of an authentic source document**

Digital Signatures and Preservation

- Digital signatures are great tools for ensuring authenticity of documents **across space …**
- **…** but not **across time**!
- Digital signatures are subject to obsolescence, and thus, only compound the problem of preservation
- Most memory institutions have announced they will not attempt to maintain encrypted or digitally signed documents transferred to them

InterPARES Trust

Preferred Means of Authentication

A chain of legitimate custody is ground for inferring authenticity and authenticate a document.

Digital chain of custody: the information preserved about the document and its changes that shows specific data was in a particular state at a given date and time.

A declaration made by an expert who bases it on the **trustworthiness of the system hosting the document and the procedures and processes** controlling its use

InterPARES Trust

Involuntary Secondary Permanence… Where?

By this time in my lecture I am certain we all have lost track of all the copies produced for distribution, redundancy, authentication, or preservation, thereby ensuring their unplanned or

involuntary secondary permanence

- All the above works like a charm with in-house digital applications and systems, because this involuntary permanence only clogs and slows down our systems, but,
- if we do not have the documents in our physical custody—say, we keep them in the cloud, can we ascertain the trustworthiness of the existing copies and can they be controlled in any degree?
- We are investigating this question in our new phase of InterPARES, called **InterPARES Trust or Itrust**

InterPARES Trust

InterPARES Trust (2013-18)

The **goal of InterPARES Trust** is to generate the theoretical and methodological **frameworks** that will support the development of integrated and consistent local, national and international **networks of policies, procedures, regulations, standards and legislation concerning digital documents entrusted to the Internet**, to ensure public trust grounded on a persistent digital memory.

InterPARES Trust is funded by a 5-year SSHRC Partnership grant and matching funds from the University of British Columbia and all the partners (who are in 6 continents and 35 countries)

Documentary Memory in the Cloud

Archives and libraries are the **trusted custodians** of our historical documentary memory. Yet, they are beginning to store their holdings in the Cloud because:

- Many of the materials they wish to preserve already **exist** in the Cloud
- **Access** is possible from any location to anyone who can use a browser
- A trusted digital repository satisfying ISO standards as well as basic preservation requirements is not **affordable** by everyone and is often inadequate to the challenge
- The **knowledge** to deal with documents produced by complex technologies is not commonly available among information professionals and is very expensive
- Strong **protection** measures are often thought to be preservation measures
- but mostly because these institutions are confronted with…

InterPARES Trust

A Generational Change

Generation Y or Millennials -- Post-1981

- integration of private and public
- *produsing*, co-authoring, crowdsourcing
- co-owning, sharing
- working from home (distributed workforce) or under a BYOD policy using multiple clouds
- media convergence
- constant connectivity
- visual language
- “liquid communication,” instantaneous impact
- ephemera

Characteristics of their Digital Output

- Documents are produced to be **viewed differently** based on choice of browser, application, and user preferences, or perceived preferences, thus the authoritative version is impossible to establish
- **Metadata may be constructed by any number of parties to manipulate** the behavior of retrieval systems that use it, rather than to describe the documents or other digital objects
- When the **goal is communication**, documents may exist in as many separate clouds as needed and may be scheduled to self-destruct (DSTRUX), or may never be destroyed **(involuntary permanence due to lassitude)**
- When the **goal is memory**, digital archives may be constructed on hard drives with copies of materials from a variety of provenances
- The digital documents of persons, organizations and institutions may be mingled and **undistinguishable**
- Three primary challenges for archival appraisal and acquisition

InterPARES Trust

Challenge 1: Reuse

While any research can be considered reuse, today

- Reuse is often *remix*, a practice which results in **derivative works** that substantively change the intent and context of the appropriated material (**we no longer have copies** but new material).

Social norms are emerging through

- **successive cycles of use and reuse,**
- **modification, repurposing, and take-down notices** (because people upload anything, rather asking for forgiveness than for permission).

Challenge 2: Sharing

- Social media platforms facilitate the **movement of material from one circle of people to another**, crossing the public-private lines and creating innumerable copies impossible to track down.
- Groups of **employees collectively create bodies of interlinked material** related to work projects (e.g. gcpedia), or common interests
- Groups assemble the **stories of activities and events using copies of documents from various provenances**, and change them overtime.
- Contributions to social media by people, programs, committees, or agencies now **dead** are **linked to ongoing, active contributions of the living, disappear, or appear as created by their successors**.
- It seems that we are moving from no original, but multiple copies to a very different scenario: **Are we confronted with an infinite number of originals?**

InterPARES Trust

Challenge 3: Location Independence

A fundamental issue with keeping documents in the Cloud today is the distinction between

- the **entity responsible for their preservation and accessibility** (e.g. the memory institution) and
- the **entity storing them** (the Provider),

and the possibility that the **jurisdiction** under which either exists is different from that in which the documents physically reside, that disposition rules are only applied within one jurisdiction, that what is destroyed are only the links to the documents (e.g. where the right to be forgotten applies)

Involuntary permanence is rising to a whole other level!

InterPARES Trust

What is the Cloud?

Often the Internet is referred to as the Cloud. Technically this is a misuse of terms.

Budapest Convention on Cybercrime, 2001

Internet providers are "entities providing users the **ability to communicate** through a computer system **that processes or stores computer data** on behalf of such communication or users." There are three "actions" related to the definition of provider: **communication, data processing** and **data storage.**

National Institute of Standards and Technology

"Cloud computing is a model for enabling convenient, on-demand network access to a **shared pool** of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction."

InterPARES Trust

Issues in the Cloud

- Transparency
- Ownership, authorship, creatorship
- Retention and disposition
- Loss of context and therefore identity
- Authenticity, reliability, accuracy: trustworthiness
- Jurisdiction
- Ethics

More Than Traditional tools

- Model policies/procedures
- Training
- Model agreements – TOS and SLA
- Leadership from national/international organizations
- Understanding of organizational culture

But we need much more, because many (or any) copy can only replace one original by having the authority and effectiveness of an original.

So, stay tuned!

www.interparestrust.org
www.ciscra.org

Director, Luciana Duranti

luciana.duranti@ubc.ca