Managing and Archiving Information and Data

Luciana Duranti &
Jean François Blanchette

InterPARES Project &
School of Library, Archival and Information Studies,
University of British Columbia

Modern Office

- Hybrid documentary systems
- Digital environments that support the manipulation and repurposing of data
- Obsolescence of systems and media
- Proprietary and idiosyncratic nature of applications

One System

- Printing out is an impediment to the workflow in the office
- Many digital documents are not printable
- Digitization is expensive in the long term
- Courts' decisions have been against routine reproduction

Consequences

- Records are less
  - reliable (manipulability),
  - retrievable (incongruence of classifications),
  - accessible (incompatibility),
  - readable or intelligible (obsolescence)
- It is difficult to prove the authenticity
- It is difficult to maintain accountability
- It is difficult to provide for long-term preservation of authentic records

How to deal with this situation

- Developing a records policy/strategy and procedure that address separately records and the other digital objects
- Focusing any such policy/strategy on the continuing reliability, accuracy and authenticity of records and data
- Recognizing that preservation of authentic electronic records and data is a continuous process that begins at the moment of creation and whose purpose is to transmit authentic information across time and space

Record

Any document created (i.e., made or received and set aside for further action or reference) by a physical or juridical person in the course of a practical activity as an instrument and by-product of it.

Records versus data

- All records are documents
- Document = recorded information
- Information = aggregation of data intended for communication over time or space
- Data = the smallest indivisible meaningful fact

Electronic Record

A record created (i.e., made or received and set aside for action or reference) in electronic form

Identifiable Characteristics of an Electronic Record

- Fixed form (i.e. its binary content is stored so that it remains complete and unaltered, and its message can be rendered with the same documentary form it had when first set aside)
- Unchangeable content
- Explicit linkages to other records within or outside the digital system through a classification code or other unique identifier
- Identifiable administrative context
- Author
- Addressee
- Writer
- Participant in or supporting an action either procedurally or as part of the decision making process

Other Characteristics

- The relation between a record and a file can be one-to-one, one-to-many, many-to-one, or many to many
- The same presentation of a record can be created by a variety of digital presentations and viceversa, from one digital presentation a variety of record presentations can derive
- It is possible to change the way in which a record is contained in a file without changing the record

What if we have scientific « data »?

- Scientists are typically less concerned with the issue of **accountability**
- **Re-use** and **access** to data is paramount:
  - where data collection is expensive (e.g., high-energy physics)
  - time-consuming (e.g., atmospheric sciences)
  - distributed (e.g., environmental science),
- Because of high volumes and high-technicality of preservation, centralization of means is often necessary

Reliability

- The trustworthiness of the record as a statement of facts or as content.
- Trust in the accuracy of the data.

Authenticity

- Refers to the fact that a record is what it purports to be and has not been tampered with or otherwise corrupted.
- Authenticity is the trustworthiness of the record as a record. To verify it, one must verify the identity and integrity of a record.
- Authenticity of data is also relative to their identity and integrity

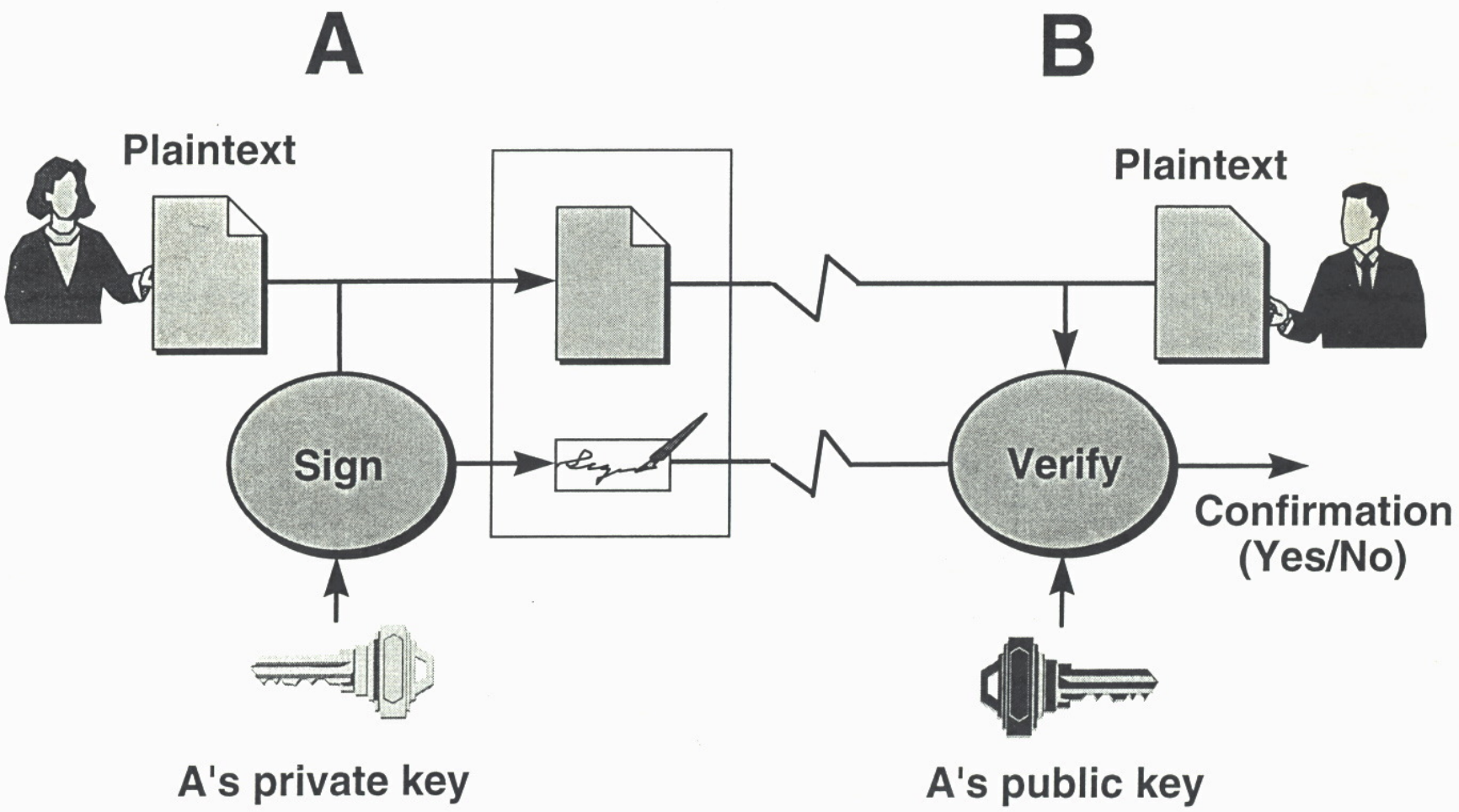
Authentication

- A declaration of authenticity, resulting either by the insertion or the addition of an element or a statement to a record, and the rules governing it are established by legislation.
- A means of proving that a record is what it purports to be at a given moment in time.

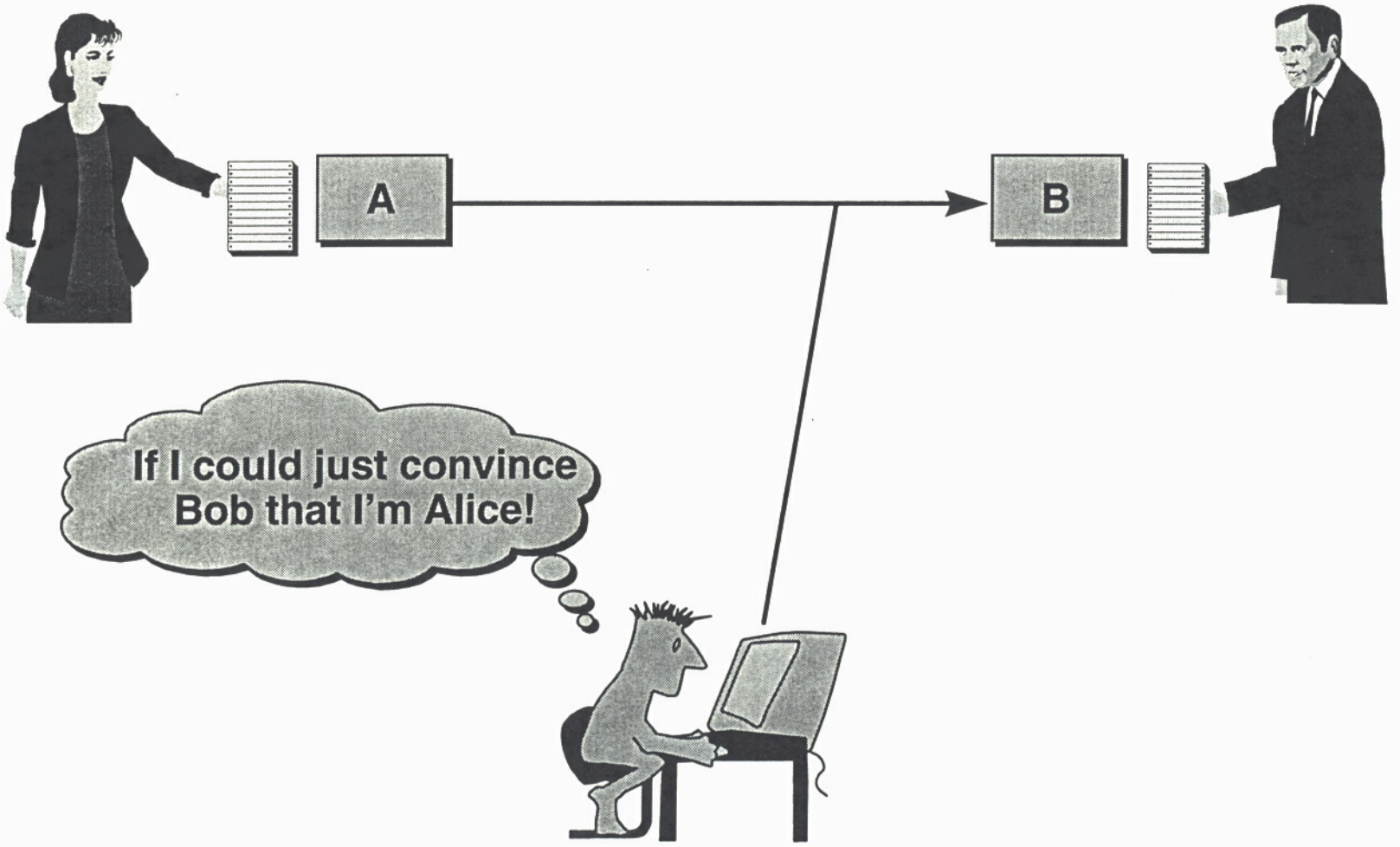
Authenticity vs. authentication

- Certain mathematical techniques are said to provide **incontrovertible** mechanisms for ensuring authenticity of digital objects (e.g., cryptographic digital signatures)
- Such technologies have been given legal value (e.g., European Directive on electronic signatures) or regulatory (SEC and hash functions.
- Digital signatures are enabled through complex and costly public-key infrastructures (PKI)
- While digital signatures are based on the same mathematical techniques as encryption, they do **not** provide confidentiality

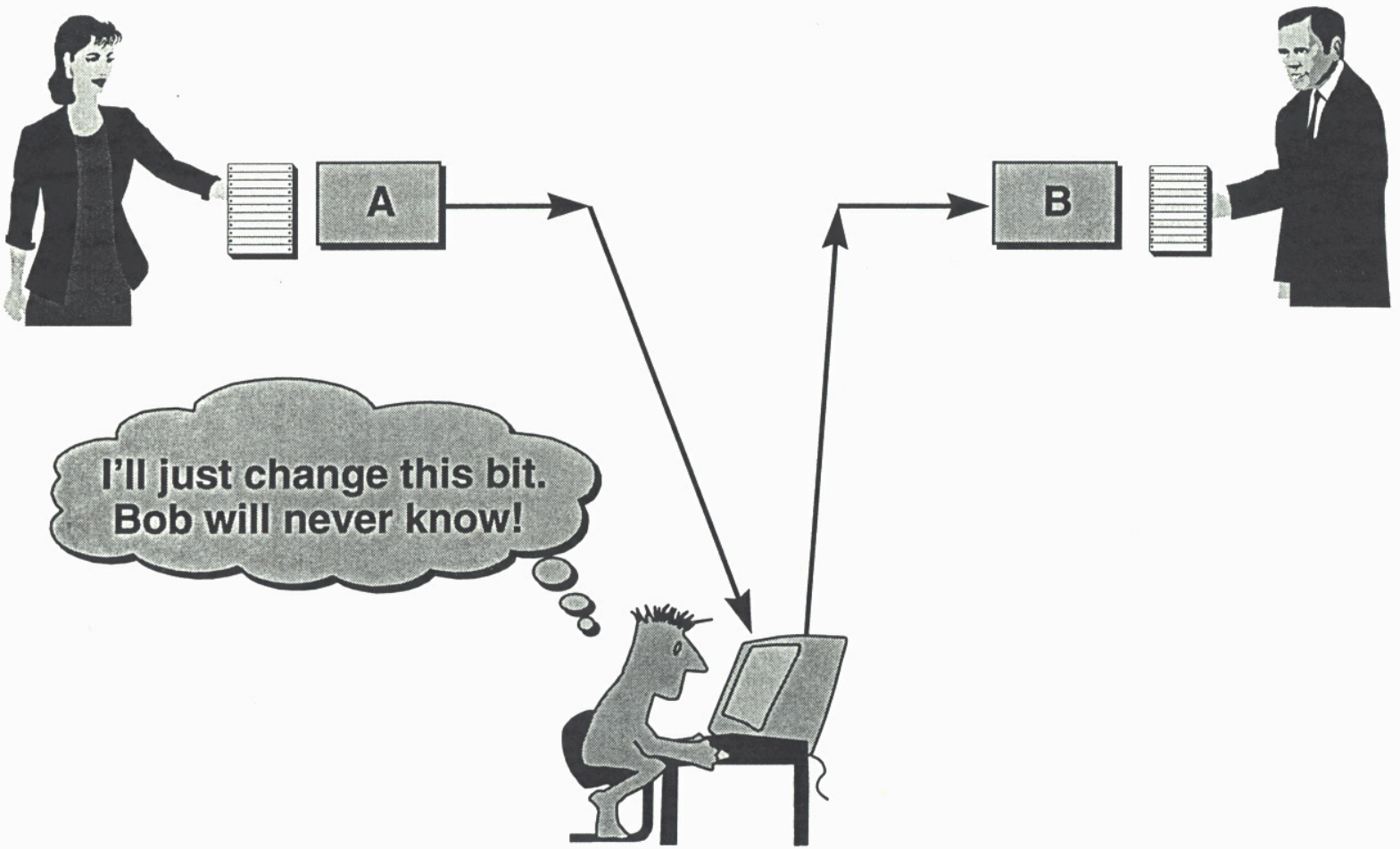
Digital Signature



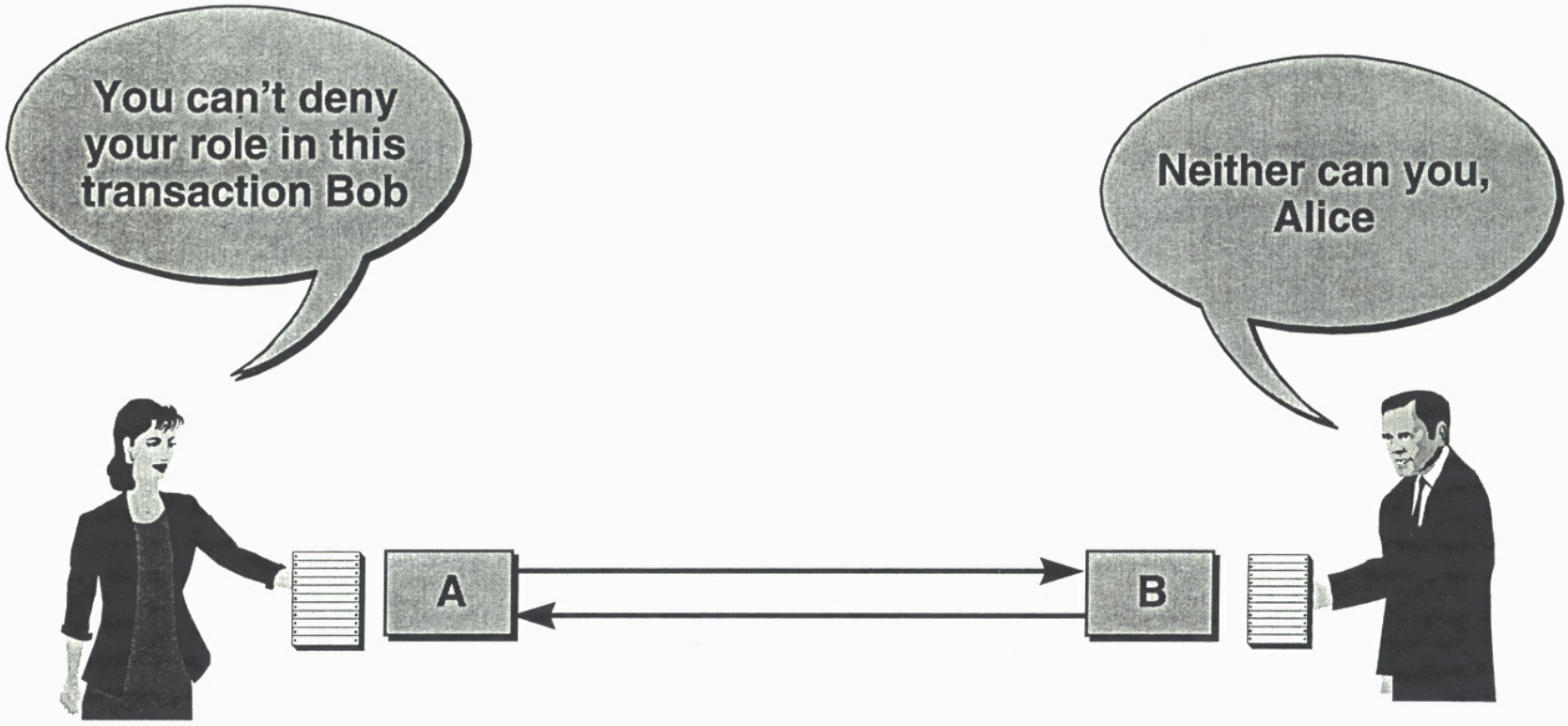
Authentication



Integrity



Non-Repudiation



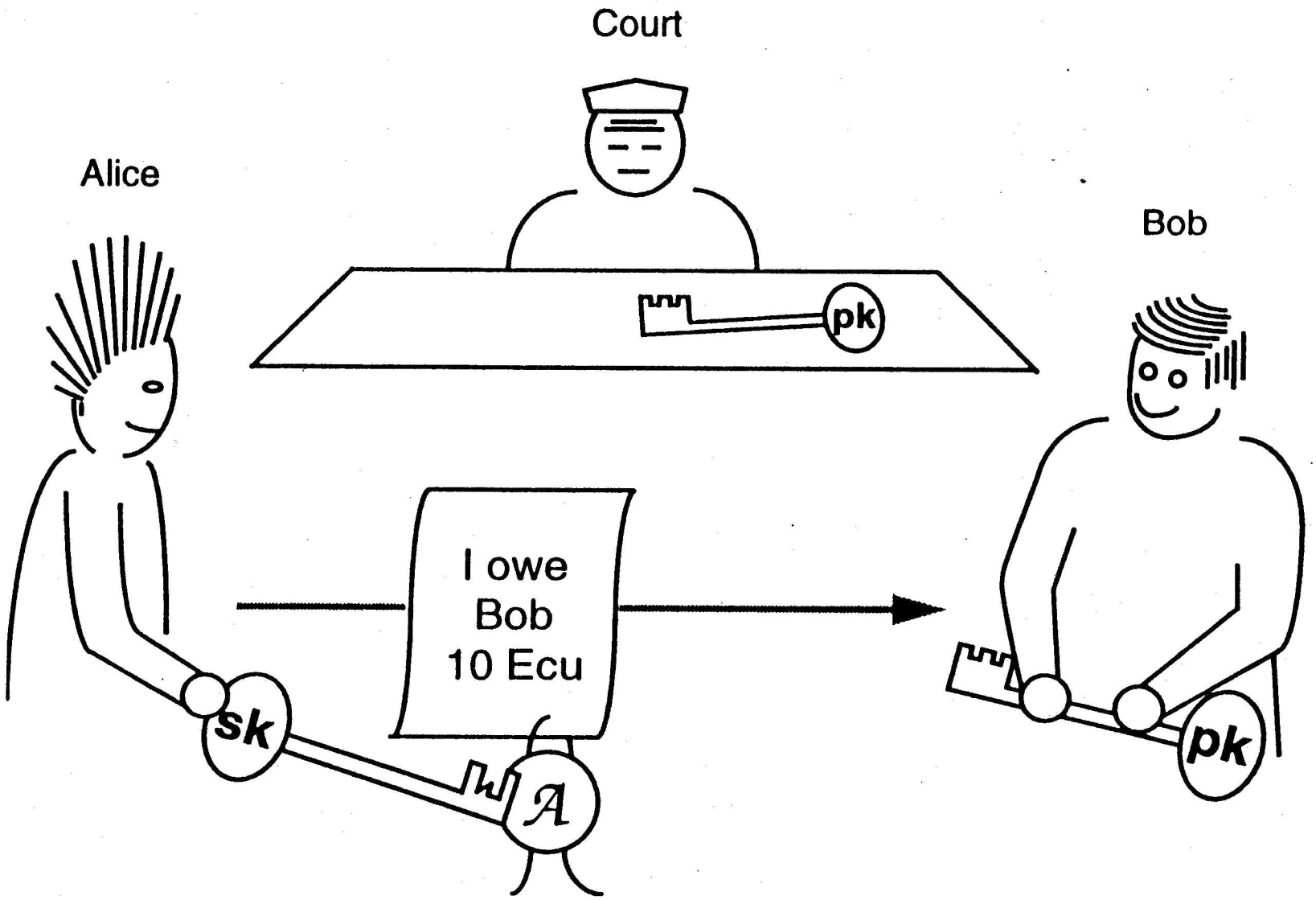


Figure 2.1. Components of ordinary digital signature schemes.
In the example, Alice has chosen a key pair (*sk*, *pk*) and published *pk*. In particular, Bob and th court know *pk*. Now, only Alice can sign her message with *sk*. Bob and the court accept a messag as signed by Alice if and only if it passes the test with *pk*.

Digital signatures and preservation

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

Threats to authenticity

Authenticity is most at risk when records are transmitted **across space** (that is, when sent between persons, systems, or applications) **or time** (that is, either when they are stored offline, or when the hardware or software used to process, communicate, or maintain them is upgraded or replaced)

Conceptual Framework for Authenticity

- In archival theory and jurisprudence, records (or data) that are relied upon by their creator in the usual and ordinary course of business are presumed authentic
- In electronic systems, the presumption of authenticity must be supported by **evidence** that a record (or data) is what it purports to be and has not been modified or corrupted in essential respects.
- To assess the authenticity of a record (or data), the preserver must be able to **establish its identity** and **demonstrate its integrity**

Identity of a Record/Data

- It refers to the attributes of a record/data that uniquely characterize it and distinguish it from other records/data. These attributes include: the names of the persons concurring in its formation (I.e., author, addressee, writer and originator); its date(s) of creation and transmission; an indication of the matter or action in which it participates; classification code or other unique identifier; as well as an indication of any attachment(s).
- These attributes may be explicitly expressed in an element of the record, in metadata related to the record/data, or implicit in its various contexts (documentary, procedural, technological, provenancial, or juridical-administrative).

Integrity of a Record

- Its wholeness and soundness. A record has integrity if it is intact and uncorrupted
- A record is intact and uncorrupted if the message that it is meant to communicate in order to achieve its purpose is unaltered
- Data are intact and uncorrupted if they are as accurate as they were when generated
- A record's or data's physical integrity, such as the proper number of bit strings, may be compromised, provided that the content and its required elements of form remain the same
- Integrity may be demonstrated by evidence found on the face of a record, in metadata related to a record/data, or in one or more of its contexts

Inference of trustworthiness

- It derives from the fact that the creator treats its records/data as trustworthy by relying on them for action or reference in the usual and ordinary course of business. It is not supportable when the records/data are no longer actively used because the motivation to maintain them accurate and authentic is less compelling.
- It is not supportable if the records/data are kept in the applications in which they are created and not subject to proper controls and if their maintenance is not continuously monitored.

Hence

- It is essential to ensure that the electronic records and data are clearly identifiable and of demonstrable integrity and that accidental corruption or purposeful tampering have not occurred since their creation.
- How do we do so?

How do we do so?

- Maintaining the records and the data in a **trusted record keeping system**
- Understanding that it is **not** possible to preserve an electronic record/data as a stored physical object: it is only possible to preserve the ability to reproduce the record/data
- Ensuring that the reproduction process is the responsibility of a **trusted custodian** having the authority and the capacity of documenting it thoroughly

Trusted Recordkeeping System

Expression of Attributes

The first requirement of a trusted recordkeeping system is that it is capable of controlling all the records/data of the creator, regardless of their physical form. This control takes place by expressing the **attributes** of each record/data and its linkage to other records/data through a register, a record profile, or a topic map.

These attributes can be distinguished into categories, the first concerning the identity of records, and the second concerning the integrity of records.

Attributes for the identity of the record/data

- Names of the persons concurring to the formation of the record, that is: name of author, writer, originator, and addressee
- Name of action or matter
- Date(s) of creation and transmission, that is: chronological date, received date, archival date, transmission date(s)
- Expression of documentary context (classification/unique identifier)
- Indication of attachments

Attributes for the integrity of the record

- Name of handling office/person
- Name of office of primary responsibility
- Indication of types of annotations added to the record
- Indication of technical modifications
- Disposition

Integrity of scientific data

- Standardized data files (with information relative to instruments, calibration, etc.)
  - In standardized numerical formats (IEEE)
  - Images files (JPEG, TIFF)
- Textual documents describing datasets, collection process and context, etc.
  - Documents (e.g., administrative records, reports) written by firms, labs, or relevant agencies
  - Major publications resulting from analysis of the dataset ("the scientific record")
  - In PDF, XML, etc.
- Cataloguing data

Trusted Recordkeeping System
Controlled management of all records/data

- Profile System
- Integrated Classification/Cataloguing and Retention System
- Controlled Disposition System

Trusted Recordkeeping System
Access Privileges

- The creator has defined and effectively implemented access privileges concerning the creation, modification, annotation, relocation, and destruction of records/data
- Maintains an audit trail of access to the records systems to control the administration and use of access privileges

Trusted Recordkeeping System

Protective Procedures: Loss and Corruption of Records

- The creator has established and implemented procedures to prevent, discover, and correct loss or corruption of records/data
- Maintains an audit trail of every transmission within the recordkeeping system
- Ensures regular system backup

Trusted Recordkeeping System

Protective Procedures: Media and Technology

The creator has established and implemented procedures to guarantee the continuing identity and integrity of records/data against media deterioration and across technological change, such as regular migration, microfilming, etc.

Trusted Recordkeeping System

Establishment of Documentary Forms

The creator has established the documentary forms of records associated with each procedure either according to legal and/or organizational requirements or its own

Trusted Recordkeeping System

Authentication of Records

If authentication is required by the juridical system or the needs of the organization, the creator has established specific rules regarding which records must be authenticated, by whom, and the means of authentication

Trusted Recordkeeping System

Identification of Authoritative Record

If multiple copies of the same record exist, the creator has established procedures that identify which record is authoritative

Trusted Recordkeeping System

Removal and Transfer of Relevant Documentation

If there is a transition of records/data from active status to semi-active and inactive status, which involves the removal of records/data from the electronic system, the creator has established and implemented procedures determining what documentation has to be removed and transferred to the preserver along with the records/data

In practice …

- The principles have been captured in various standards, e.g.,
  - DoD 5015.2: US federal government vendors
  - SEC Rule 17a-4: any financial body
  - FDA 21 CFR part 11: pharmaceutical industry
- Software vendors have developed products which complies to various degrees with such standards and regulation:
  - DoD: 5015.2: TRIM (Tower Software)
  - SEC Rule 17a-4: Centera (EMC)
  - MoREQ: R/KYV (Valid Information Systems)

TRIM

- Care of documents in their entire continuum and strong classification features, which enable:
  - (a) providing linkages between individual records;
  - (b) records are named in a consistent manner over time;
  - (c) assisting in the retrieval of all records relating to a particular activity;
  - (d) appropriate retention periods for record;
  - (e) security protection appropriate for sets of records;
  - (f) allocating user permissions for access to or action on particular groups of records
- Capturing the initial context and supporting additional context relationships as they evolve greatly facilitates information management and retrieval and inferences of authenticity

EMC Centera

- Designed to meet SEC regulations, which
  - requires that the electronic storage media preserve the records exclusively in a non-rewriteable and non-erasable format.
  - does not include storage systems that only mitigate the risk a record will be overwritten or erased, i.e., software applications to protect electronic records, such as authentication and approval policies, passwords or other extrinsic security controls.
- Based on “fixed-content addressing“ — using cryptographic hash functions, a unique fingerprint is calculated from each document
- Provides mathematical assurance that documents are never modified, even if using rewritable media (hard disks)
- Can be used as a back-end to a RMS like TRIM.

Smart Enterprise Suite approach

Single-vendor bundles of integrated modules including:

- Content management
- Collaboration tools
- Multi-channel access
- Information retrieval
- Expertise location
- Process management
- Portal framework

Alternatives to SES

Record, Document, Information Management System (RDIMS: www.rdims.com and www.rdims.gc.ca):

- core document and records management products from Hummingbird
- document management, imaging, basic workflow and records management modules from Open Text and Documentum
- reporting tools from Crystal Decisions

Alternatives to SES

Coquitlam Enterprise Document Management System (CEDMS):

Built from the merging of a Document Management System and a Record Management System.

Includes a file plan imported using SQL script written by Concerta, an ACCESS database, and a dozen file formats.

Digital preservation

- The transmission of digital information objects across technological boundaries (computer platforms, operating systems, applications) created by technological obsolescence
- A digital object possesses:
  - A physical dimension, as an inscription on a physical carrier (punch card, mag. tape, optical disc)
  - A logical dimension, as this inscription must be recognized and processed by software
  - A conceptual dimension, as an object produced and to be understood within a specific context

Thus …

- In order to preserve a digital object, we must be able to identify and retrieve all of its digital components, i.e., the logical and physical objects necessary to **reconstitute** the conceptual object
- That is, to access any digital object, **stored** bit sequences must be **interpreted** as logical objects and **presented** as conceptual objects
- In the paper-and-ink world, the basis of preservation is the caring for the integrity of the physical carrier itself, but…

Reproducing digital objects

- Digital preservation is not a simple process of preserving physical objects (stored bit sequences), but one of **preserving the ability to reproduce the objects,** and this process is complete only when the objects are successfully output!
- Preserving a digital object **does not imply** preserving its physical and logical components and their relationships without alteration!

Migration

- The most prevalent preservation strategy is currently that of **migration**:
  - Migrate **media** as technology changes, using state-of-the-art technology for **storage** and **access**
  - Migrate **data formats** as technology changes, using state-of-the-art technology for **output**
- Migrating data formats involves changes to the object's bitstream, i.e., **authenticity ≠ integrity of bitstream**!

Criteria for reproduction

- The problem becomes, “Which changes are permissible and/or beneficial?”
- Given that a digital information object is something that can only be **re-constructed** by using software to process stored inscriptions, it is necessary to have an **explicit model or standard** that provides a **criteria for assessing the authenticity** of the re-constructed object
- Given that the migration process offers the opportunity of corrupting digital objects without leaving any trace and does change aspects of the objects, it is necessary that it be carried out by a **trusted custodian**

Trusted Custodian

Controls over Records/Data Transfer, Maintenance, and Reproduction

The procedures and system(s) used for maintaining the records in the long term and reproducing them must embody adequate and effective controls to guarantee the records' identity and integrity, and specifically that:

- unbroken custody is maintained;
- security and control procedures are implemented and monitored;
- the content remains unchanged after reproduction

Trusted Custodian

Documentation of the Reproduction Process and its Effects

The activity of reproduction must be documented, and this documentation should include:

- the date of the reproduction and the name of the responsible person;
- the relationship between the digital objects acquired from the creator and the copies produced by the preserver;
- the impact of the reproduction process on their form, content, accessibility and use; and
- in those cases where a copy of a record is known not to fully and faithfully reproduce the elements expressing its identity and integrity, such information has been documented by the preserver, and this documentation is readily accessible to the user

Key points concerning preservation

- Technology cannot determine the solution to the long-term preservation of electronic records
- Organizational needs define the problem and archival principles must establish the correctness and adequacy of each technical solution
- Solutions to the preservation problem are inherently dynamic

Key points concerning preservation (2)

- Preserving records and data over the long term is a separate and complex function
- Managers (in scientific research and elsewhere) have yet to take the full measure of the specificity and consequences of this activity
- Through ignorance, lack or responsibility and misunderstanding of this issue, numerous digital records will be lost in the coming years
- Hopefully, these losses may serve as triggers for a collective awakening to this coming crisis

For more information

- InterPARES site: http: www.interpares.org
- Very good digital preservation site: http://www.nla.gov.au/padi/
- Questions, comments:
  - luciana@interchange.ubc.ca
  - Jean-Francois.Blanchette@ubc.ca