MINISTERIO DE CULTURA

XXX SEMINARIO
Sistema Nacional de Archivos

SITA2022
Semana de Innovación Tecnológica en Archivos

GESTIÓN DEL GOBIERNO ABIERTO Y PRESERVACIÓN DIGITAL: REFLEXIONES DESDE LOS ARCHIVOS

NOV 8-11

|EVENTO ACADÉMICO VIRTUAL|

SNA
Sistema Nacional de Archivos

ARCHIVO GENERAL DE LA NACIÓN COLOMBIA

MINISTERIO DE CULTURA

Digital Preservation - the findings of InterPARES

Thursday, November 10, 2022
11 am GMT -5

Corinne Rogers, MAS, PhD

Project Coordinator, InterPARES Trust AI
University of British Columbia
corinne.rogers@ubc.ca
www.interparestrustai.org

InterPARES Trust AI

GESTIÓN DEL GOBIERNO ABIERTO
Y PRESERVACIÓN DIGITAL:
REFLEXIONES
DESDE LOS ARCHIVOS
SNA
Sistema Nacional de Archivos
ARCHIVO GENERAL DE LA NACIÓN COLOMBIA

Agenda

- 25 years (almost) of InterPARES through 5 phases
- Researching issues of trustworthy creation and preservation of digital records across generations of technological evolution
- Case studies in government digital recordkeeping and preservation
- Conclusions

InterPARES – 5 phases

InterPARES 1 (1998-2001) www.interpares.org

- Researched issues pertaining to digital records in databases and office management systems in the course of administrative activity
- Focused on developing theory and methods to ensure preservation of authenticity using archival diplomatics
- Studied records from the perspective of the records preserver

InterPARES – 5 phases

InterPARES 2 (2002-2007) www.interpares.org

- Researched issues pertaining to digital records in dynamic and interactive systems in artistic, scientific, and government activity
- Examined issues of authenticity, reliability, and accuracy over the lifecycle
- Studied records from the perspective of the records creator

InterPARES – 5 phases

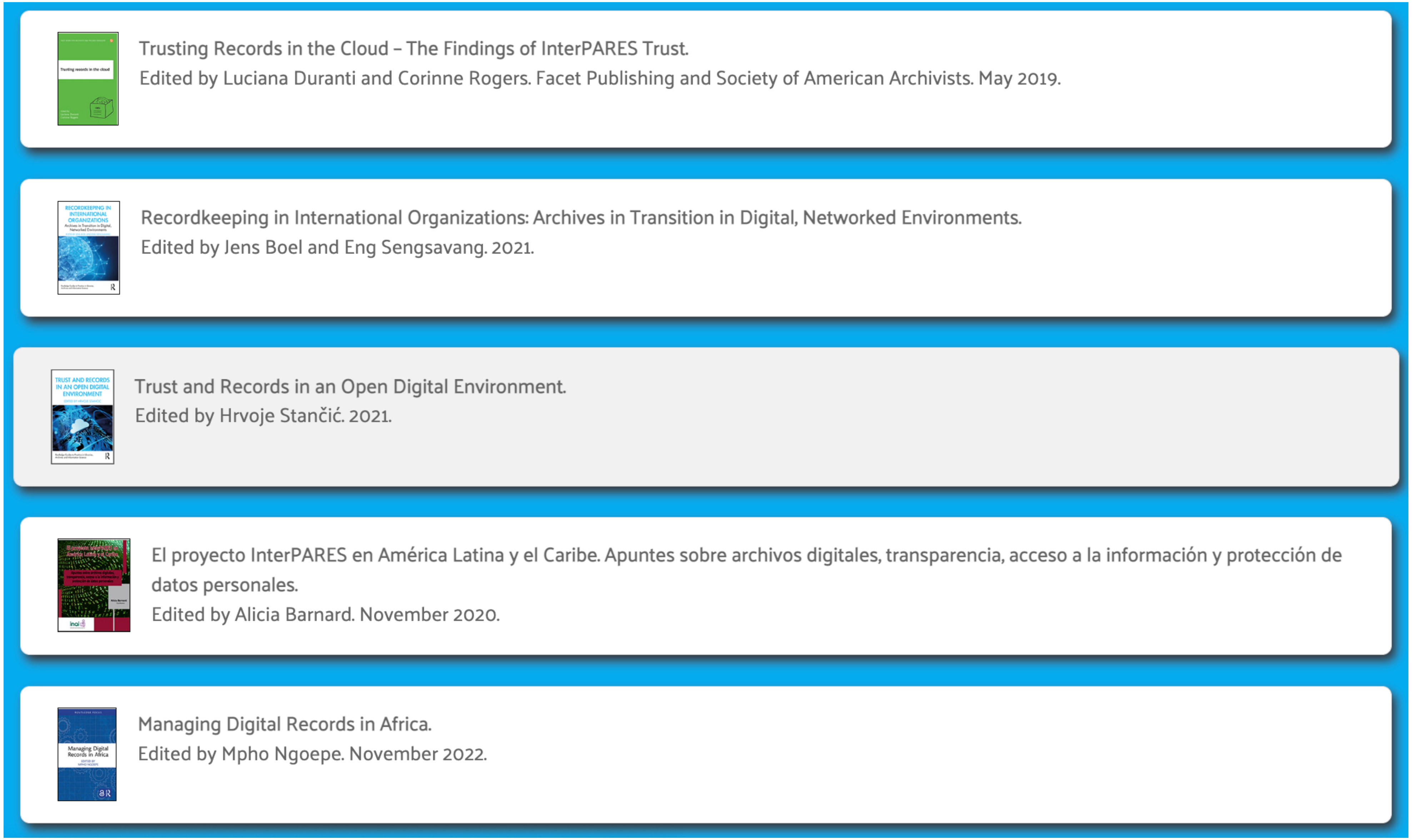
InterPARES 3 (2007-2012) www.interpares.org

- Put theory into practice in archives / records units in organizations with limited financial or human resources
- Applied and tested the findings of InterPARES 1 and 2 to implement sound programs supporting the creation and preservation of digital records that could be shown to be authentic, reliable, accurate

InterPARES – 5 Phases

InterPARES Trust (2013-2019) www.interparestrust.org

- Researched the impact of always-on, networked communications technologies and cloud computing services on records management & recordkeeping, maintaining trustworthy records & supporting client/citizen perception of trustworthiness of records, in order to
- Ensure public trust grounded on evidence of good governance, strong digital economy, & persistent digital memory



Trusting Records in the Cloud – The Findings of InterPARES Trust.
Edited by Luciana Duranti and Corinne Rogers. Facet Publishing and Society of American Archivists. May 2019.

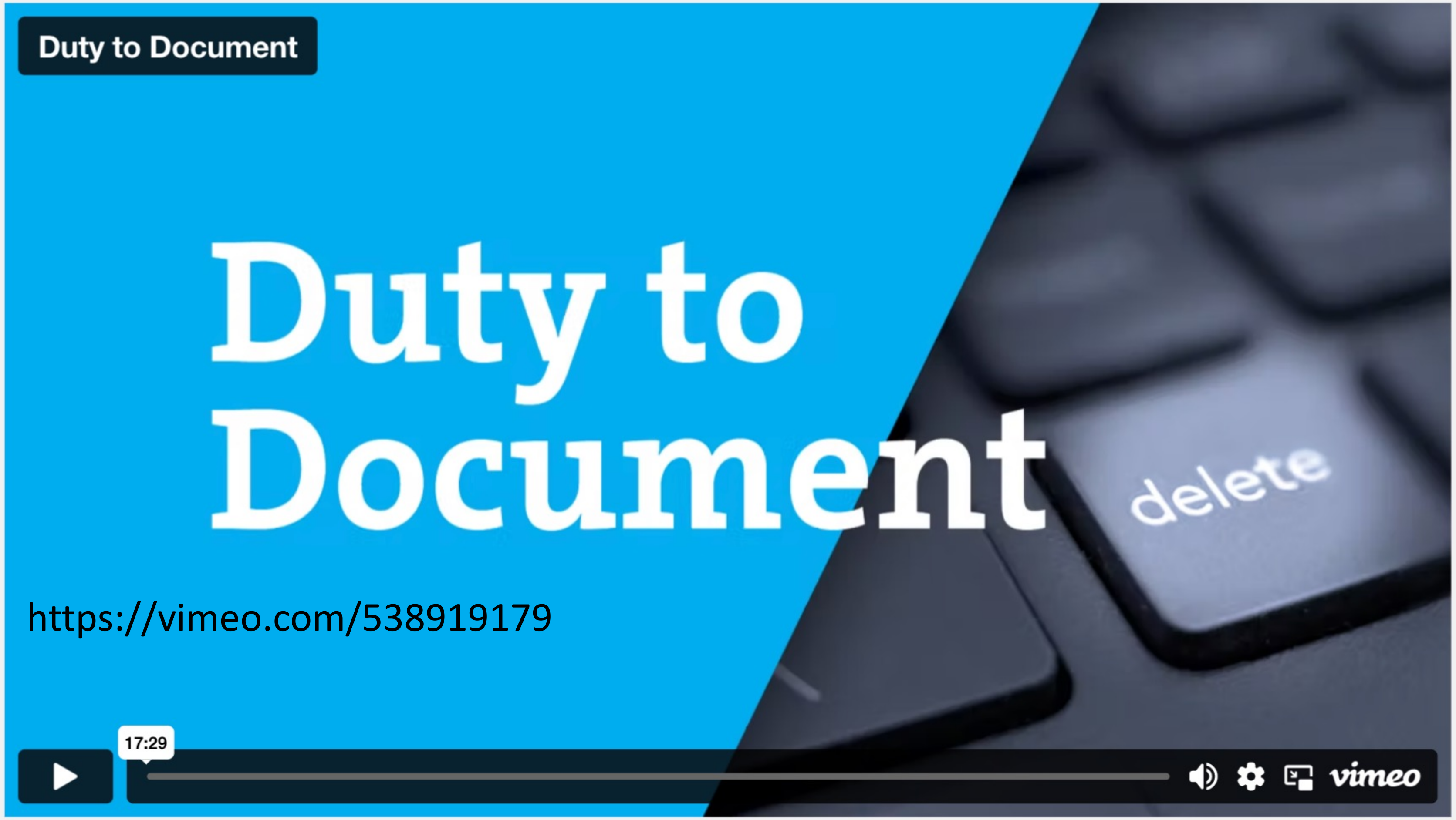
Recordkeeping in International Organizations: Archives in Transition in Digital, Networked Environments.
Edited by Jens Boel and Eng Sengsavang. 2021.

Trust and Records in an Open Digital Environment.
Edited by Hrvoje Stančić. 2021.

El proyecto InterPARES en América Latina y el Caribe. Apuntes sobre archivos digitales, transparencia, acceso a la información y protección de datos personales.
Edited by Alicia Barnard. November 2020.

Managing Digital Records in Africa.
Edited by Mpho Ngoepe. November 2022.

This documentary was produced by InterPARES Trust and the Canadian Institute for Information and Privacy Studies. It looks behind the scenes at the techniques that politicians and bureaucrats use to avoid accountability.



https://vimeo.com/538919179

InterPARES – 5 Phases

InterPARES Trust AI (2021-2026) http://www.interparestrustai.org/

- Researching the use of Artificial Intelligence to support the ongoing availability and accessibility of trustworthy public records
  - Identifying specific AI technologies that can address critical records and archives challenges;
  - Determining the benefits and risks of using AI technologies on records and archives;
  - Ensuring that archival concepts and principles inform the development of responsible AI; and
  - Validating outcomes from Objective 3 through case studies and demonstrations.

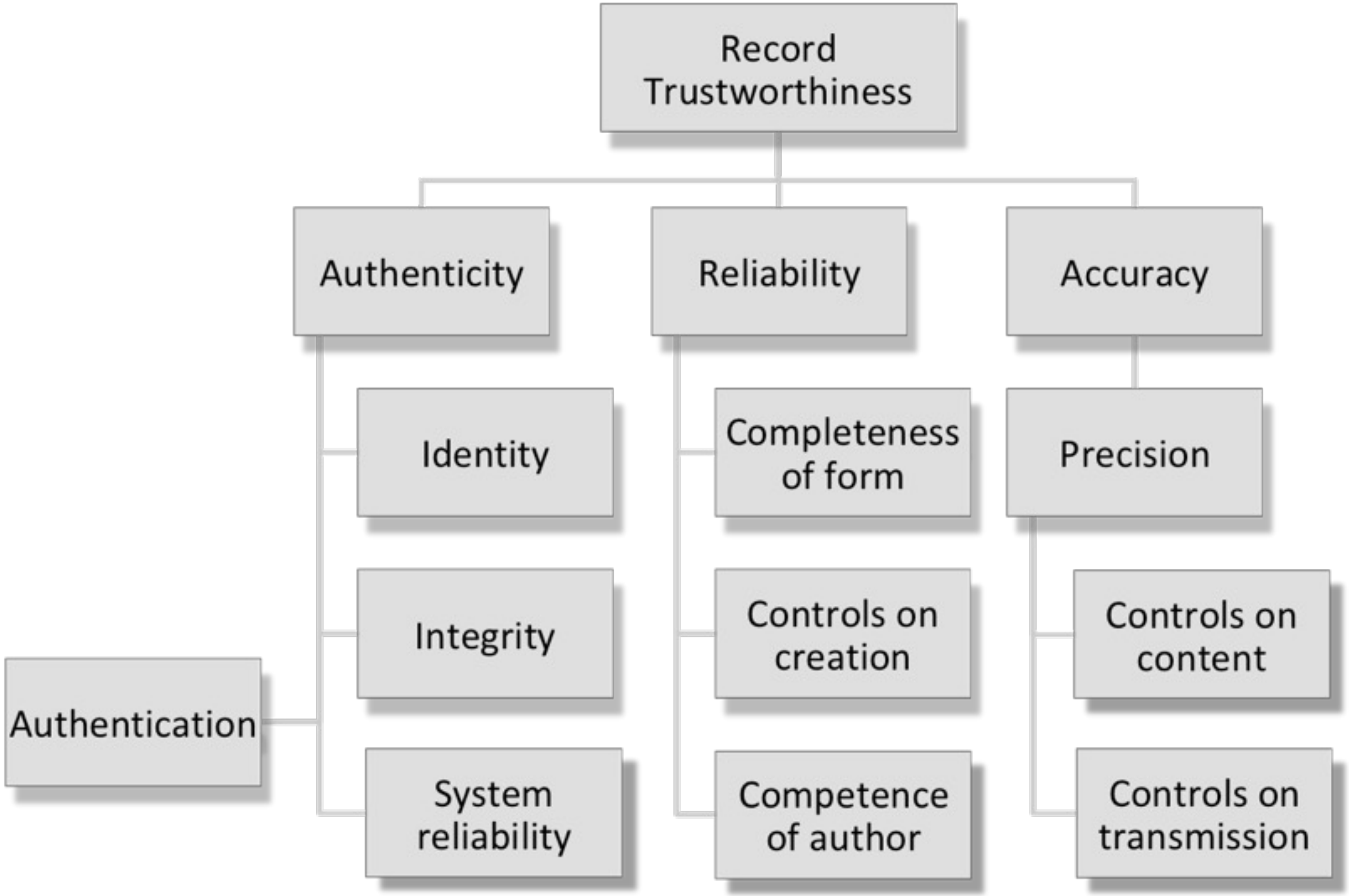
Social Sciences and Humanities Research Council of Canada
Conseil de recherches en sciences humaines du Canada
Canada

Impact since inception

- Created a worldwide network of researchers and scholars; trained a generation of archivists
- Legislation: Italy, China
- Standards: DOD 5015.2 (2007), MoReq 2 (2008), OAIS (2009), CGSB 72.34 (2017), MPEG ISO
- Policies & procedures: all participating countries, public/private sector
- Curricula for continuing education & university training
  - ICA Education Modules for Digital Preservation (2012 with translation to Chinese, Spanish, Arabic)
  - Summer school in San Benedetto, Italy (2023)
  - Curricula in partner universities on five continents

Foundations – IP1, 2

Foundations of InterPARES Research – Archival Diplomatics



- Record Trustworthiness
  - Authenticity
    - Identity
    - Integrity
    - System reliability
    - Authentication
  - Reliability
    - Completeness of form
    - Controls on creation
    - Competence of author
  - Accuracy
    - Precision
      - Controls on content
      - Controls on transmission

InterPARES 1 – Preservation Task Force

- If information is intended to serve as a record of an action or state of affairs, the message it transmits must be **fixed**
- Fixity of digital records is at risk because of changes in the way a digital record is represented in storage and the way it is presented for use
- Saving and accessing digital records entails transforming the content, structure, and appearance of the record
- Integrity of the record depends on guaranteeing that no alteration or degradation of content takes place between the stored and presented record

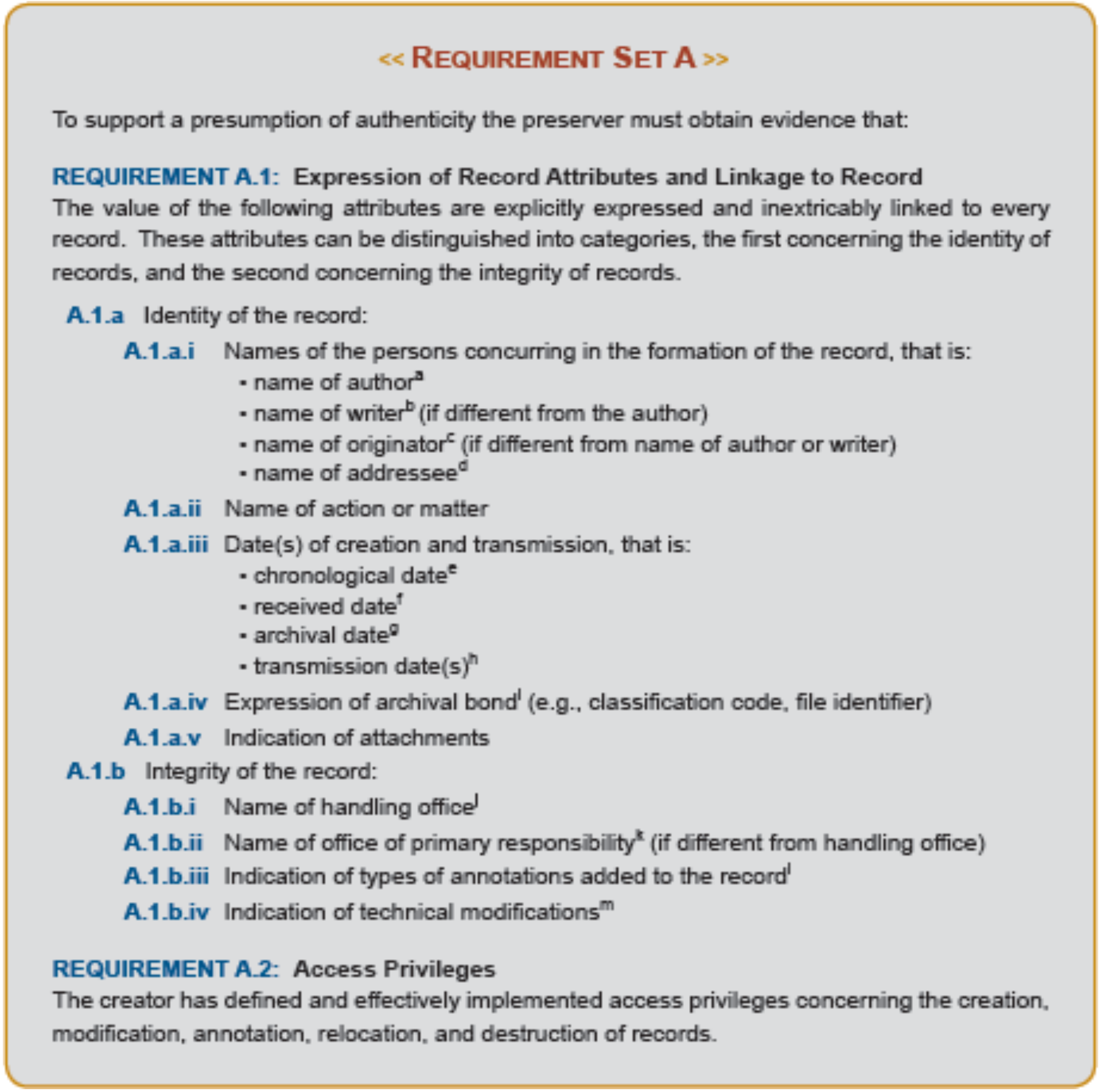
InterPARES 1 – Preservation Task Force

- Preservation of digital records extends over the entire life cycle from creation to disposition and reproduction
- A digital record cannot be said to have been preserved unless it can be **reproduced in authentic form** – that is, its identity is clear and its integrity proven
- Both its intellectual form and its digital components must be preserved
- **The ability to preserve digital records begins at creation**
- The traditional concept of **unbroken chain of legitimate custody** to protect and prove authenticity **is no longer enough**

InterPARES 1 – Preservation Task Force

- We require knowledge of a **Chain of Preservation**
- This extends chain of custody to include information about the records creator’s practices to support a presumption of authenticity, and
- Information about the processes of reproducing the records
- These requirements are outlined in
  - **Benchmark Requirements Supporting the Authenticity of Digital Records**
  - **Baseline Requirements Supporting the Production of Authentic Copies**

Benchmark Requirements



<< REQUIREMENT SET A >>

To support a presumption of authenticity the preserver must obtain evidence that:

REQUIREMENT A.1: Expression of Record Attributes and Linkage to Record
The value of the following attributes are explicitly expressed and inextricably linked to every record. These attributes can be distinguished into categories, the first concerning the identity of records, and the second concerning the integrity of records.

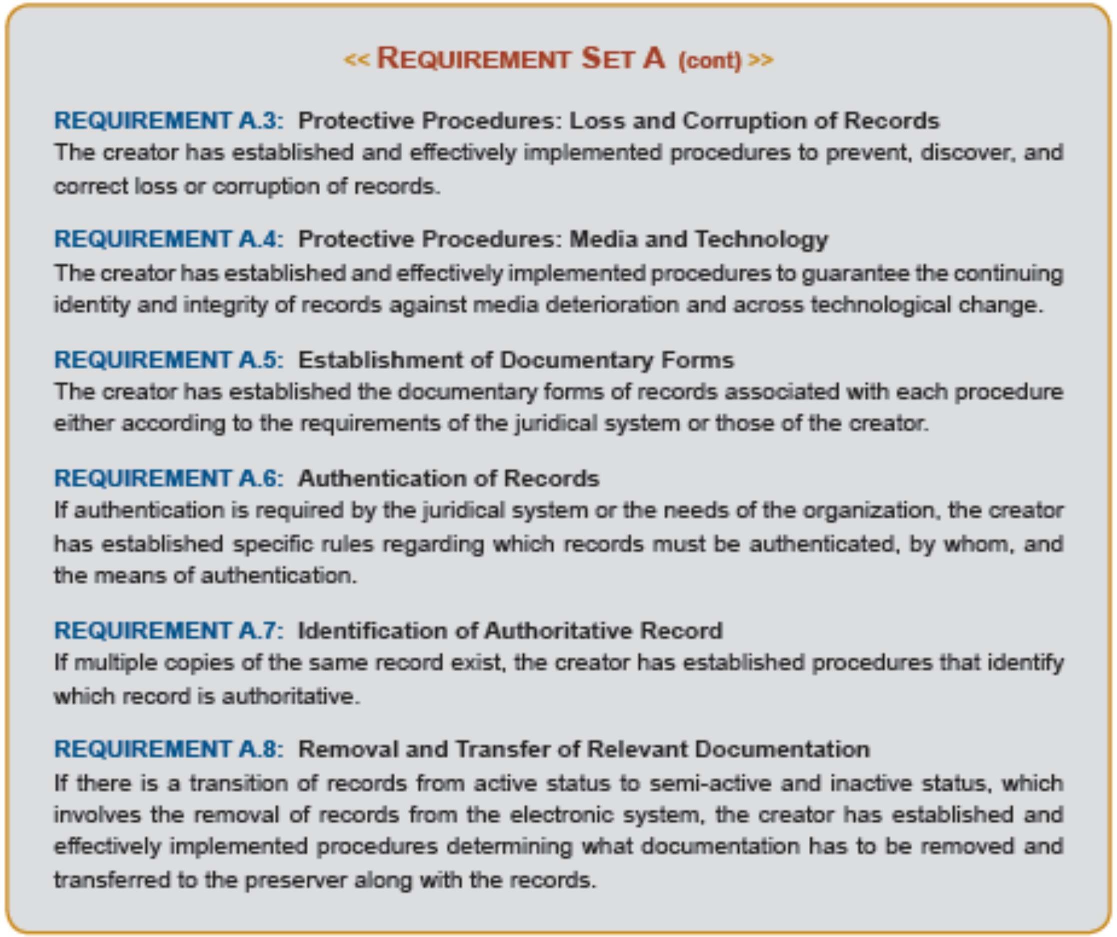
A.1.a Identity of the record:

- A.1.a.i Names of the persons concurring in the formation of the record, that is:
  - name of author[a]
  - name of writer[b] (if different from the author)
  - name of originator[c] (if different from name of author or writer)
  - name of addressee[d]
- A.1.a.ii Name of action or matter
- A.1.a.iii Date(s) of creation and transmission, that is:
  - chronological date[e]
  - received date[f]
  - archival date[g]
  - transmission date(s)[h]
- A.1.a.iv Expression of archival bond[i] (e.g., classification code, file identifier)
- A.1.a.v Indication of attachments

A.1.b Integrity of the record:

- A.1.b.i Name of handling office[j]
- A.1.b.ii Name of office of primary responsibility[k] (if different from handling office)
- A.1.b.iii Indication of types of annotations added to the record[l]
- A.1.b.iv Indication of technical modifications[m]

REQUIREMENT A.2: Access Privileges
The creator has defined and effectively implemented access privileges concerning the creation, modification, annotation, relocation, and destruction of records.



<< REQUIREMENT SET A (cont) >>

REQUIREMENT A.3: Protective Procedures: Loss and Corruption of Records
The creator has established and effectively implemented procedures to prevent, discover, and correct loss or corruption of records.

REQUIREMENT A.4: Protective Procedures: Media and Technology
The creator has established and effectively implemented procedures to guarantee the continuing identity and integrity of records against media deterioration and across technological change.

REQUIREMENT A.5: Establishment of Documentary Forms
The creator has established the documentary forms of records associated with each procedure either according to the requirements of the juridical system or those of the creator.

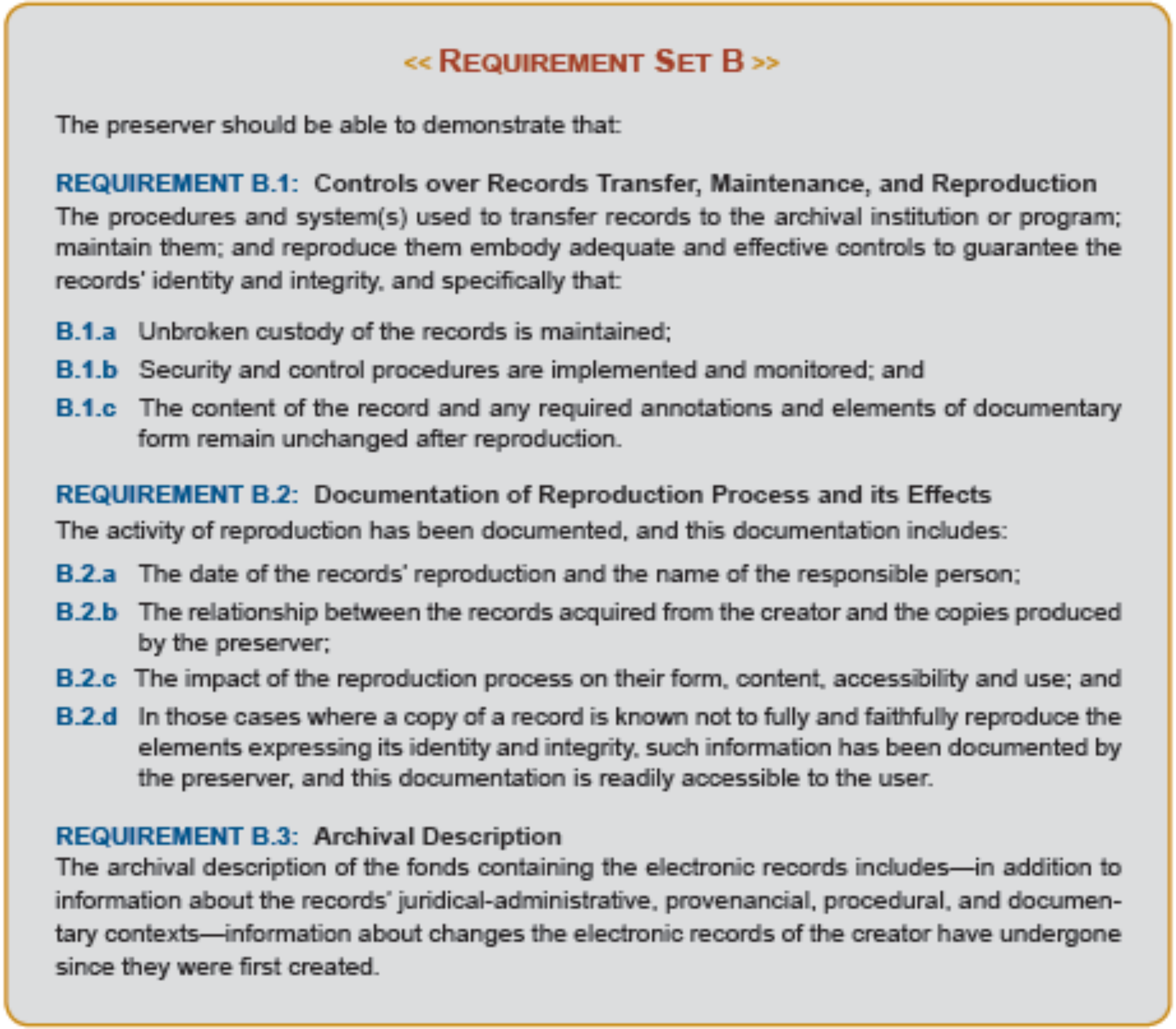
REQUIREMENT A.6: Authentication of Records
If authentication is required by the juridical system or the needs of the organization, the creator has established specific rules regarding which records must be authenticated, by whom, and the means of authentication.

REQUIREMENT A.7: Identification of Authoritative Record
If multiple copies of the same record exist, the creator has established procedures that identify which record is authoritative.

REQUIREMENT A.8: Removal and Transfer of Relevant Documentation
If there is a transition of records from active status to semi-active and inactive status, which involves the removal of records from the electronic system, the creator has established and effectively implemented procedures determining what documentation has to be removed and transferred to the preserver along with the records.

Supporting the Presumption of Authenticity of Digital Records

Baseline Requirements Supporting the Production of Authentic Copies



<< Requirement Set B >>

The preserver should be able to demonstrate that:

REQUIREMENT B.1: Controls over Records Transfer, Maintenance, and Reproduction
The procedures and system(s) used to transfer records to the archival institution or program; maintain them; and reproduce them embody adequate and effective controls to guarantee the records' identity and integrity, and specifically that:

B.1.a Unbroken custody of the records is maintained;

B.1.b Security and control procedures are implemented and monitored; and

B.1.c The content of the record and any required annotations and elements of documentary form remain unchanged after reproduction.

REQUIREMENT B.2: Documentation of Reproduction Process and its Effects
The activity of reproduction has been documented, and this documentation includes:

B.2.a The date of the records' reproduction and the name of the responsible person;

B.2.b The relationship between the records acquired from the creator and the copies produced by the preserver;

B.2.c The impact of the reproduction process on their form, content, accessibility and use; and

B.2.d 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.

REQUIREMENT B.3: Archival Description
The archival description of the fonds containing the electronic records includes—in addition to information about the records' juridical-administrative, provenancial, procedural, and documentary contexts—information about changes the electronic records of the creator have undergone since they were first created.

Case studies in government records

InterPARES 2 – City of Vancouver VanMap (2005)

- VanMap is a web-based GIS system maintained by the IT department
- Its main purpose is to meet the needs of internal users in providing services to Vancouver's citizens and businesses, and it is accessible to all City staff through the City's intranet
- It supports core functions – zoning, development permit approval, business license approval, emergency planning, water and sewer, traffic control, etc.
- A subset of data is made directly available to the public via the City's website

InterPARES 2 – City of Vancouver VanMap

- Is VanMap a record, or does it produce records?
- VanMap data are continuously updated by various City departments, other government agencies and utility companies – end-user interaction if dynamic and experiential
- Some data are overwritten without being saved
- Data are viewed as maps – these views are not saved – there is no definable archival bond
- The procedural contexts for using VanMap in operations is not documented
- VanMap can be considered to create and contain **potential records**

InterPARES 2 – City of Vancouver VanMap

- To turn these *potential* records into records:
    - Take a complete image (snapshot) that is live and fully functional at the end of each business day
    - Schedule removal of these snapshots at set intervals and keep them as fixed, stable records for that time period
    - Develop detailed descriptions of the business processes in which VanMap is involved and how VanMap is used in order to understand the relationship between the records of business processes and Van Map

InterPARES 2 – City of Vancouver VanMap

- To preserve these *potential* records as records:
  - Clearly identify what the record is
  - Preserve the ability to render the data as interactive maps
  - Decide what presentation elements (e.g. colours, fonts) need to be preserved

InterPARES 3 – City of Surrey Drive Migration (2008)

- City of Surrey was (in 2008) the 12th largest city in Canada; 2nd largest city in British Columbia
- The City held millions of information assets on large servers with multiple drive paths; many were low value records duplicated across drives, existing in many versions, and had met their legal/operational needs; others were mission critical
- The goal was to design, construct, and deliver a records system to manage unstructured digital records throughout the information lifecycle;
- To appraise existing unstructured digital records;
- To ingest into the repository, or move offline for authorized deletion; and
- To provide a sustainable foundation for e-business standardization, workflow integration, enterprise-wide collaboration, and paper reduction, in a business environment characterized by continued, rapid growth

InterPARES 3 – City of Surrey Drive Migration

Status at the outset of the project:

- 15 years of legacy, unstructured digital records
- Drive space provided by IT department
- Ad hoc folder management, no standardization
- No digital records management:
  - No record classification
  - No automated version or duplication control
- No preservation strategy
- No authorized disposal

InterPARES 3 – City of Surrey Drive Migration

The new system should provide:

- An enterprise-wide single point of access
- Improved customer service and information sharing
- Built-in digital records management processes
- Ensure legal compliance and reduce records-related risk
- Provide audit trails
- Increase ability to respond to citizen’s requests and legal discovery
- Increase accountability
- Improve vital records protection and disaster recovery
- Ensure long-term viability and preservation of unstructured digital records

InterPARES 3 – City of Surrey Drive Migration

ECM (Enterprise Content Management) components:

- Program management
- Records & information governance
- InfoBank requirements specification (the City’s information bank)
- Support technology
- InfoBank Design & Build
- **Drive migration**
- InfoBank Implementation

InterPARES 3 – City of Surrey Drive Migration

Drive migration mission statement:

To develop a process whereby business units in the City of Surrey can prepare existing unstructured electronic records held in shared drives for appraisal and ingest into the InfoBank repository for operational use or long-term preservation, or offline authorized deletion.

InterPARES 3 – City of Surrey Drive Migration

InterPARES worked with City of Surrey staff to deliver:

- Migration Assessment: an inventory of assets and requirements
- Migration Methodology: identified and customizied best practices for appraisal and migration
- Migration Tools: Requirements for migration tools
- Migration Manual: A communications manual describing migration procedures and best practices for business units
- These were based on the Chain of Preservation model, Benchmark and Baseline Requirements for Authentic Records; Policy Requirements

InterPARES 3 – City of Surrey Drive Migration

Drive Migration Toolkit

- A roadmap for migrating from shared network drives to an enterprise-wide records management and preservation system
  1. Defining the corporate context
  2. Defining the technical context
  3. Defining the business context
  4. Completing a business appraisal
  5. Completing a technical appraisal
  6. Completing a records appraisal
  7. Migrating records
  8. Appendices

InterPARES Trust – Records in the Cloud

- What is the impact of always-on, networked communications technologies and cloud computing services on records management & recordkeeping, maintaining trustworthy records & supporting client/citizen perception of trustworthiness of records?

- Cloud service provider contracts checklist (NA14)
- Managing Records of Citizen engagement (NA08)

Trust & Chain of Custody

- We keep records (sometimes over long periods of time) as evidence of activity, and as memory of action, & to prove accountability – we must trust them
- In archival terms, we trust records based on proof of records' authenticity, reliability, & accuracy
- In legal terms, trust is expressed through rules of admissibility of documentary evidence (common law systems)
- Demonstrable chain of responsible custody is key to both

Challenges & Ethical Duties

- The challenges most frequently discussed when considering cloud services represent present concerns with current data (**data-centric thinking**):
  - Is data secure from alteration or interference?
  - Can personal privacy be protected?
  - Can regulations and laws be observed in the face of cross-jurisdictional data transfer?
  - What guarantees of continuity of service exist?
  - How will data breaches be handled?

Recordkeeping Challenges

- Recordkeeping challenges look beyond the immediate present, reaching into the past, and projecting into the future (**record-centric thinking**)
  - Can context of records be protected?
  - Can provenance be demonstrated?
  - Can retention & disposition be carried out?
  - Can access and usability be assured over time?
  - Can intellectual rights be respected?

Ethical Issues

- Privacy and confidentiality
- Security
- Access & the digital divide
- Intellectual rights
- Jurisdiction
- Identity, de-identification, re-identification
- Data analysis (sentiment, emotion), risk of profiling

Trustworthy records and record systems

- Records are deemed trustworthy based on an assessment of their authenticity, reliability, and accuracy
- These records requirements depend on trustworthy, controlled systems
- Do cloud services meet the standard of trustworthy records systems?

Tools for evaluation of cloud services

- InterPARES Trust developed several tools to help records and archives professionals address records and data issues, regardless of the degree of cloud adoption
- These tools help evaluate the benefits and risks from the perspective of *recordkeeping* based on *archival science*

Cloud Service Provider contracts as instruments of trust: Purpose & Research question (NA14)

- This study explored the contract – specifically the contract between a client and a cloud service provider – as a tool for building trust
- How effectively do cloud service contracts meet the needs of records managers, archivists, and information governance professionals?

Selected contracts

- Boilerplate contracts & documents
  - Terms of Service (ToS)
  - Service Level Agreements (SLA)
  - Privacy policies, Acceptable Use policies, Security terms,
- Jurisdictions
  - Canada, United States, Europe

Amazon.com (USA); Bluelock (USA); Dropbox (USA); Egnyte (USA); GoGrid (USA); **Google (USA);** ProfitBricks (USA); Rackspace (USA); CityNetwork (Sweden); SAP (Belgium); Pathway Communications (Canada)

Contracts review

- Findings:
  - Several legal documents exist
    - Terms of Service
    - Service Level Agreements
    - Privacy Policies
    - Acceptable Use Policies
  - Little standardization of terms
  - “Often incomprehensible to majority of users”
  - Wide-ranging exclusions of liability favor the providers
  - Terms may change

Comparative Analysis

- Regardless of jurisdiction, sector, or industry, common risks to records exist:
  - Unauthorized access
  - Privacy breach
  - Loss of access, control
  - Lack of transparency of service
  - Lack of ability to negotiate service
  - Location ambiguity
  - Contract ambiguity

Records considerations often not adequately addressed

- Data ownership
- Availability, retrieval and use
- Data storage and archival preservation
- Data retention and disposition
- Security, confidentiality, privacy
- Data location and cross-border data flow
- End of service; contract termination

CSP Checklist - sections

- Agreement
- Data Ownership and Use
- Availability, Retrieval, and Use
- Data Storage and Preservation
- Data Retention and Disposition
- Security, Confidentiality, and Privacy
- Data Localization and Cross-border Data Flows
- End of Service; Contract Termination

Open government: Government-Citizen Engagement (GCE) Relationship (NA08)

- Researched the implications of open government, open data, and big data on the management of digital records in an online environment
- Developed a guide to
  - Enhance awareness of the relationship between recordkeeping and GCE initiatives
  - Suggest approaches for addressing recordkeeping issues that impact trust relationships between governments and their citizens

Analysis and guidance from the perspective of records, recordkeeping

The Guide addresses issues and suggests strategies at the level of:

- Policy
- Governance and Management
- People and personnel
- Standards and practices
- Technology
- Awareness and shared understanding

Plan for managing records of GCE

- Analyze GCE in your organization
- Identify records-related issues from a business perspective (implications for success of GCE)
- Identify strategies relevant to address these issues
- Identify all the individuals who will be concerned or affected by the issues
- Discuss the issues with relevant individuals
- Prepare plan to bring forward for approval

InterPARES Trust AI

- As in previous phases, ITrust AI focuses on
  - maintaining the trustworthiness of digital records overtime
  - digital means of trustworthy access to and preservation of records in all media and form
- The assessment of the authenticity of digital material is always an inference based on extrinsic elements such as significant properties included in identity and integrity metadata, and relies on circumstantial evidence such as
  - the integrity of the system hosting it at any given moment in time,
  - the policies and procedures controlling such system, and
  - the technology encrypting the record or securing the access to it.
- Can we use Artificial Intelligence to verify authenticity?

About the Project

- The studies are international and interdisciplinary
- The focus is the use of existing, and development of new, AI tools to assist in all aspects of archival functions:
  - Creation and use of trustworthy records
  - Appraisal and acquisition of archival material
  - Arrangement and description
  - Retention and preservation
  - Management and administration of records and archives
  - Reference and access

A taste of the work so far (1st Symposium, October 27, 2022, Canary Islands)

- Tutorials on Natural Language Processing, Part of Speech Tagging, Named Entity Recognition, Text Translation, Speech Processing
- Digital Preservation and AI – Critical Challenges, Hrvoje Stancic, University of Zagreb
- Opportunities and Challenges for AI-Assisted Digitization of Cultural Heritage Materials, Eng Sengsavang, UNESCO Paris
- From Natural Language Processing to Appearance Based Approaches: Challenges and Opportunities for Archival Science, Emanuele Frontoni, University of Macerata, Italy

A taste of the work so far (1st Symposium, October 27, 2022, Canary Islands)

- Archival Values and AI, Jim Suderman, Toronto, Canada
- Knowledge Representation and Breaking Down Silos at the Bank of Canada, Alex Richmond & Marielle Saint-Germain, Bank of Canada
- The Elusiveness of AI-based Automated Records Classification, Umi Mokhtar, University of Kebangsaan, Malaysia
- Capturing and Preserving the AI Process as Paradata for Accountability and Audit Trail Purposes, Pat Franks, San Jose State University
- Challenges of Preserving the Digital Twin, Tracey Lauriault, Carleton University

Conclusions

- All phases of InterPARES have been concerned with
  - trustworthiness over time of records in digital environments
  - Digital means of trustworthy access to and preservation of records in all media and forms
- Across generations of technologies the challenges remain constant
  - Risks to authenticity from the moment of creation
  - Risks in the active environment
  - Risks in the preservation environment

Risks to authenticity

- “In analyzing the live systems, we were specifically concerned with (1) establishing the status of digital entities contained within them as records, and (2) identifying the elements of such records specifically associated with identity and integrity.” (MacNeil, Providing Grounds for Trust)
- “…few if any information systems existing in organizations create records, or at least records which are adequate to serve as *evidence* of business transactions.” (Bearman, Trant, Electronic Records Research Working Meeting)

Risks in the active environment

- Lack of guarantee of identity and integrity
- Record systems vs information systems
- Record systems: designed to capture and retain *records* for purposes of evidence and memory
  - Time-bound, inviolable
- Information systems: to store discrete information (data) that can be recombined and reused without reference to documentary context
  - Timely, manipulable
- Information systems support day-to-day effectiveness, but are a liability with respect to evidence of actions, accountability, authenticity

Risks in the preservation environment

- Unbroken chain of custody is no longer enough to ensure trust
- Ingest – insufficient metadata
- Access
- Disposition

Regardless of the technology involved…

- In digital records, **content, structure, and form are not inextricably linked**
- The record as a stored entity is distinct from its manifestation on a computer screen, and its **digital components** have to be considered in addition to its **documentary form**
- Digital records are **vulnerable** (easy to destroy, lose, corrupt, tamper with, or become inaccessible if not protected) yet **persistent** (forever there, if not purposefully destroyed)
- When we save a record, we take it apart in its digital components. When we retrieve it, we generate a copy: there are **no originals** in the digital environment
- **Hence, it is not possible to preserve digital records: we can only preserve the ability to re-produce or re-create them**
- **Digital preservation** is the process of generating and maintaining **authentic copies** of digital materials and keeping them accessible during and across different generations of technology over time, irrespective of where they are stored
- **Authenticity is the major issue when it comes to digital records**

www.interpares.org

www.interparestrust.org

www.interparestrustai.org

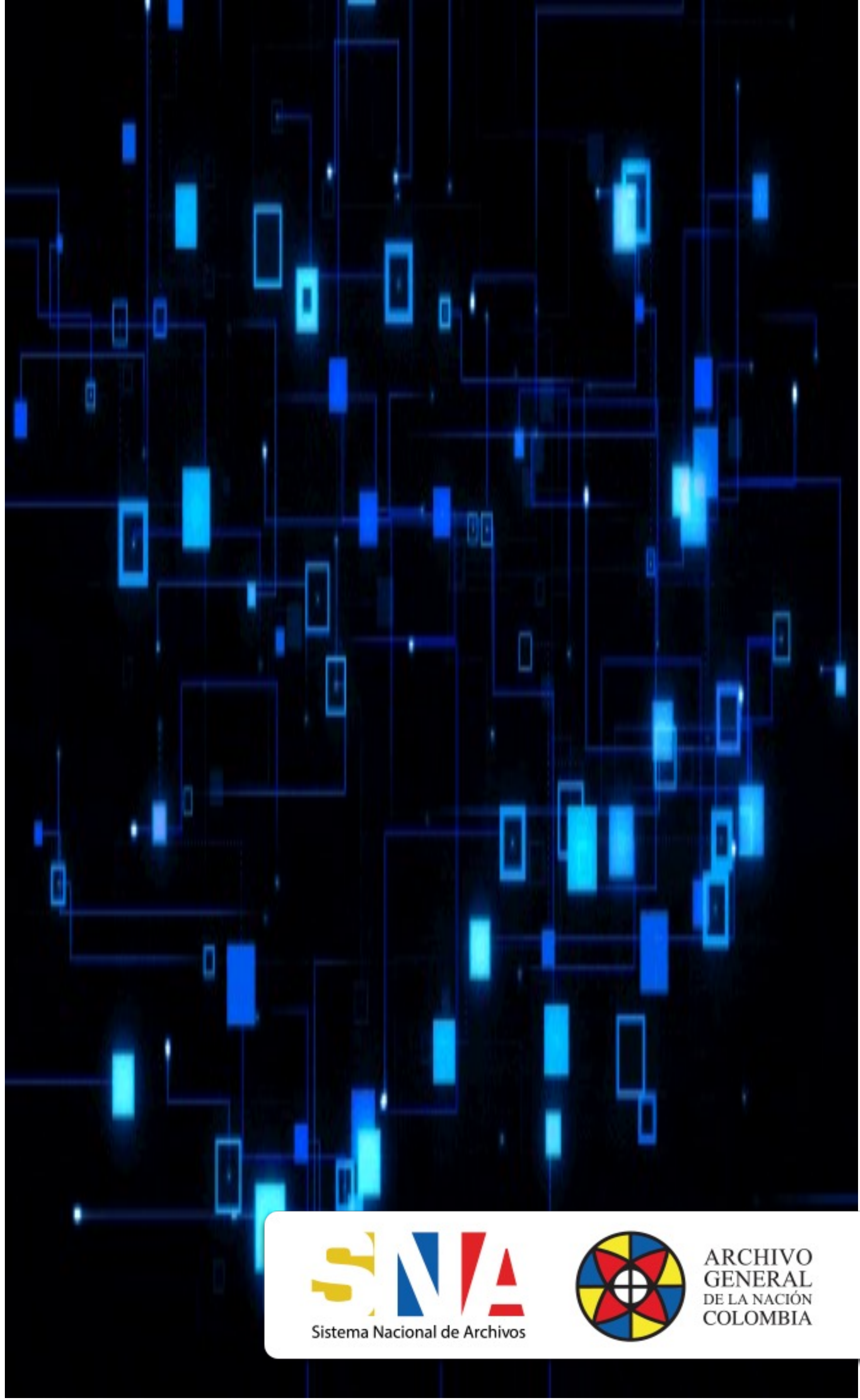
corinne.rogers@ubc.ca

Tag cloud by Ashashyou (Own work) [CC BY-SA 4.0 (http://creativecommons.org/licenses/by-sa/4.0)], via Wikimedia Commons

MINISTERIO DE CULTURA

XXX SEMINARIO
Sistema Nacional de Archivos

SITA 2022
Semana de Innovación Tecnológica en Archivos



SNA
Sistema Nacional de Archivos

ARCHIVO GENERAL DE LA NACIÓN COLOMBIA