The Long-Term Preservation of Authentic Electronic Records

Luciana Duranti

University of British Columbia
SLAIS-#831-1956 Main Mall
Vancouver, British Columbia
Canada
luciana@interchange.ubc.ca

Abstract

This paper presents the InterPARES project, its goal, objectives and domains of inquiry, its fundamental concepts and assumptions, its methods and general outcomes. It then focuses on one of its products, the conceptual requirements for the authenticity of electronic records, and concludes with a glance at the second phase of the project.

1. Introduction

The International Research on Permanent Authentic Records in Electronic Systems, known as the InterPARES project, began in 1999 and is nearing the completion of its first phase. Its goal was to develop the theoretical and methodological knowledge essential to the permanent preservation of authentic records generated and/or maintained electronically, and, on the basis of this knowledge, to formulate model policies, strategies and standards capable of ensuring that preservation. The first phase of InterPARES was set out to deal with digital records mandated for accountability and administrative needs, which are usually created in very large databases and document management systems. The creation, maintenance and use of this type of records by the organization producing them are highly controlled, thus the InterPARES research has focused on the preservation of authenticity after the records are no longer needed by the creating body.

To achieve the project goal and to address the complex variety of issues that affect the permanent preservation of authentic electronic records, the investigation was divided into four interrelated domains, each representing a research objective, supported by a dedicated interdisciplinary and multicultural task force, and including a set of research questions. The domains were: 1) conceptual requirements for the preservation of authentic electronic records; 2) appraisal criteria and methods for selecting authentic electronic records to be permanently preserved, 3) methods, rules and procedures for the permanent preservation of authentic electronic records; 4) principles that should guide the development of international strategies and standards for the long-term preservation of authentic electronic records, and criteria for developing from them national and organizational policies and strategies respecting cultural diversity and pluralism.

The research uses concepts and methods from a variety of disciplines, including diplomatics, archival science, law, computer science, computer engineering, and statistical sciences. The team includes co-investigators from the public and private sectors of Canada, United States, United Kingdom, Ireland, Sweden, Netherlands, France, Portugal, Italy, Australia, China, and Hong Kong. The intellectual mediation and integration that occur among disciplines and cultural traditions are expressed in the project's glossary of terms.

2. Fundamental Concepts and Assumptions

The work began with a definition of the fundamental concepts, in consideration of the interdisciplinariety of the project and the tendency of the disciplines involved to borrow terms from each other attaching them quite different meanings. The terms on the use of which the researchers needed to agree at the outset were "record" and "authenticity". *Record* was defined as any document made or received in the course of activity as a means and instrument for it, and set aside for action or reference. An electronic record was defined as a record maintained for use in electronic form. In order to distinguish records among all other kinds of information that may reside in a

Proceedings of the 27th VLDB Conference, Roma, Italy, 2001

digital system, the research team named several identifiable characteristics, deriving from the fact that a record can be viewed as a complex of elements and their interrelationships. First and foremost, a digital entity is a record if it has a fixed form. A record is considered to have a fixed form when its binary content, including indicators of its documentary form, is stored in a manner that ensures it remains complete and unaltered, and when the message of the record is capable of being rendered with the same documentary form it had when it was first set aside. In addition to a fixed form, a digital entity that is a record must have an unchangeable content, an explicit linkage with other records within or outside the digital system through a classification code or some other unique identifier, an identifiable administrative context, an author, an addressee and a writer. Finally, a digital entity that is a record must participate in or support an action either procedurally or as part of the decision making process.

Authenticity was defined as the trustworthiness of records as records, as distinct from reliability, which is the trustworthiness of a record as a statement of fact and the exclusive responsibility of the record creator, rather than of its preserver. An authentic record is a record that is what it purports to be, immune from corruption or tampering. Authenticity is not to be confused with *authentication*, which is a declaration of authenticity, resulting either by the insertion, the addition or the attachment of an element, a statement or a component to a record, which allows one to verify that the record is what it purports to be at that point in time. Authentication does not establish authenticity over time.

In both archival theory and jurisprudence, records that are relied upon by the creator in the usual and ordinary course of business are presumed authentic. However, digital information technology creates significant risks that electronic records may be altered, inadvertently or intentionally. Therefore, in the case of records maintained in electronic systems, the presumption of authenticity must be supported by evidence of it. In order to assess the authenticity of an electronic record, one must be able to establish its identity and to demonstrate its integrity. The *identity* of a record comprises the names of its author, addressee, writer and originator, its date, the name of the action or matter, its status of transmission (i.e. original, draft or copy), its relationship to the other records of the same creator, and the indication of attachments. Knowledge of all these attributes is essential to establish the identity of any record. The *integrity* of a record relates to its wholeness and soundness: a record has integrity when it is essentially intact and uncorrupted. This does not mean that the record must be precisely the same as it was when first created for its integrity to exist and be demonstrated. Even in the paper world, with the passage of time, are subject to deterioration, alteration and/or loss. In the electronic world, the fragility of the media, the obsolescence of technology and the idiosyncrasies of systems likewise affect the integrity of records; therefore, there is not such a thing as an uncorrupted record. **When we refer to an electronic record, we consider it essentially intact and uncorrupted if its identity is clear and the message that it is meant to communicate in order to achieve its purpose is unaltered**. This implies that its physical integrity, such as the proper number of bit strings, may be compromised, provided that the articulation of the content and its required elements of form remain the same. The attributes that constitute the identity and integrity of a record may be explicitly expressed in an element of the record or in metadata related to the record, or be implicit in its various contexts

The application of the two fundamental concepts of record and authenticity was based on two assumptions. The first is that the authenticity of records in live systems is threatened during transmission across space (i.e., person-to-person communication) and time (i.e., maintenance by the creating body for future reference), especially when this involves migration from an obsolescent to a new technology. The second is that it is not possible to preserve an electronic record, but only the ability to reproduce it. Further, it is virtually impossible to deliver any preserved electronic record in such a way that none of its elements have changed. To attest the authenticity of a preserved electronic record, then, is to demonstrate that no essential element of the record has changed. This requirement can be satisfied only if the preservation function is exercised in such a way that any changes that do occur are identified and documented. This can only be accomplished if one knows what the elements of the record were when the record was selected for preservation. After that, one faces the need to demonstrate that none of the changes affected the ability to prove the identity and integrity of the record.

3. The Work of The Task Forces: Methods and Outcomes

The task force responsible for formulating conceptual requirements for authenticity established an analytical framework for understanding existing and future records in electronic systems by developing a "Template for Analysis" according to diplomatic concepts and methods. The Template is a decomposition of an electronic record into its constituent elements: it defines each element, explains its purpose, and indicates whether, and to what extent, that element is instrumental in verifying the authenticity of the record over the long term. To populate and test the validity of the template, the task force has conducted case studies of digital systems that contain, generate, or have the potential or possibility to create electronic records. The studies include large databases used to manage, for example, student records, patent granting, securities or bank transactions; document management systems used to support agency-wide administrative functions, such as the drafting and

management of procedures, as well as specific operational functions, such as the issuing of permits for the transportation of hazardous waste or the conditional release and pardon of criminal offenders; geographic information systems, such as land data systems; and web application systems, such as trademarks systems. The instrument for conducting the case studies was a "Case Studies Interview Protocol" developed from the template and refined after each round of case studies on the basis of a statistical analysis of the data resulting from them. The whole process was guided by grounded theory, a method for discovering concepts and hypotheses and developing theory directly from the data under observation. This means that cases were selected according to their potential for helping to expand on or define the concepts or theory that had already been developed: data collection and analysis proceeded together. After including the case studies results in the "Template Data Gathering Instrument," which maps the responses to the interview questions to the elements of the Template for Analysis, a diplomatic analysis of each case study was conducted for the purpose of establishing whether the electronic systems examined contained records and, if the answer was affirmative, to determine whether the elements of the records were brought together and how, whether they manifested themselves in a way similar to traditional records, which elements the creating organization considered essential for verifying the records' authenticity, what kind of procedural controls exercised over the system and the records contained in it supported the organization's presumption of authenticity, and what type of records the system contained. From the understanding developed in the course of this work, the task force has developed conceptual requirements for the preservation of authentic electronic records that will be tested and finalized in the next few months. They will be discussed in the next section.

The task force responsible for developing appraisal criteria and methods for electronic records that respect the authenticity requirements has analysed the methods and procedures employed by archival institutions for the appraisal of electronic records and developed activity models of the appraisal function for electronic records. A major expected benefit of its work is the specification of the kinds of contextual information that needs to be gathered during appraisal. However, also all the other steps involved in conducting selection of electronic records, including timing, location, agents, manner and feasibility, are modelled, and several case studies are walked through the modelled appraisal process for the purpose of analysing the outcome.

The task force charged with identifying and developing the preservation procedures and resources required for implementing the outcomes of the first two domains has developed a formal model of the process of preserving electronic records, a template for applying the model to specific sets of records, a model of the entities that are involved in preserving electronic records, and guidelines that institutions and organizations can use to articulate comprehensive and coherent frameworks to guide the development and operation of a preservation system specifically tailored to the records each institution is responsible for preserving. Central to the entire preservation model is the concept of what it means, at an empirical level, to preserve an electronic record.

The task force responsible for developing a framework for the formulation of international standards and national and organizational policies and strategies has developed a methodology and a procedure for the distillation of principles and criteria guiding the formulation of standards, policies and strategies from the findings and final recommendations of the three other task forces. The procedure will heavily involve the national research teams. This represents the most delicate point of the research, when the universal concepts, principles and methods developed by internationally constituted task forces are brought into specific national, organizational and cultural realities and so contextualised. At this time, the task force is in the process of comparing international and national standards, as well as national and organizational policies that are relevant to the work of the task forces with their drafted findings, deliverables, and recommendations.

4. Conceptual Requirements For Authenticity

As mentioned earlier, the first task force has established benchmark requirements supporting the presumption of authenticity of electronic records maintained by their creator. The records affected by these requirements can be distinguished in two categories. The first category comprises those records that exist as created, having not undergone processing that has altered their documentary form, architecture or any part of their technological context. The second category comprises those records that result from a migration process from an obsolescent technology to a new one.

Once one has assessed the evidence of the identity and integrity of the records of the creator, one can make a presumption of their authenticity based upon how many of the requirements have been met and to what degree. The requirements are, therefore, cumulative: the higher the number of satisfied requirements, the higher the presumption of authenticity. The degree to which an individual requirement is satisfied also affects the degree of presumption. This is why these requirements are termed 'benchmark' requirements. Where there is an insufficient basis for a presumption of authenticity, a verification of authenticity will be needed. Unlike the presumption of authenticity, which is established on the basis of the requirements, this verification involves a detailed examination of the records themselves in all of their contexts. Methods of verification include, but are not

limited to, a comparison of the records in question with copies that have been preserved elsewhere or with backup tapes, textual analysis of the record's content, a study of audit trails over time, and the testimony of a trusted third party.

It is an assumption of the task force that the records are presumed or verified authentic in the appraisal process by the entity responsible for their preservation. Thus, the maintenance of their authenticity after that process is the exclusive responsibility of the preserver, who must carry forward the records by reproducing them, and authenticating the copies so produced. The production of authentic copies is a complex endeavor, which must be regulated by a second set of requirements. Unlike the benchmark requirements for authentic electronic records, all the requirements for the production of authentic copies of electronic records must be met before the preserver can attest to the authenticity of the copies in its custody. This is why this second set of requirements is termed 'baseline' requirements.

Traditionally, the official preserver of the records has been the person entrusted with issuing authentic copies of them. For a copy to be considered authentic, the preserver needed simply to attest that the copy conformed to the record being reproduced. With electronic records, the difficulties related to preservation make it prudent for the preserver to produce and maintain documentation of the activity of reproduction to support its attestation of authenticity. Thus, an electronic copy of an authentic electronic record is authentic if attested to be so by the official preserver and if such attestation is supported by the preserver's ability to demonstrate that all the requirements for the production of authentic copies have been satisfied. By virtue of this attestation, the copy is deemed to conform to the record it reproduces until proof to the contrary is shown.

The conceptual benchmark and baseline requirements apply to any type of electronic record. Among the systems analyzed as case studies, all those containing records implemented at least two of the benchmark requirements. The main concern of the research team was, however, that systems which, because of their function in the organization, are meant to contain records attesting to specific actions and transactions, such as universities' student information systems, and several government registration and inventory systems, given the fluidity of their content, did not contain records but only data, and made therefore impossible to implement the requirements. In fact, the most significant, if not unexpected, finding of the case studies was that most large databases used in electronic governance and administration are unable to serve accountability purposes, let alone to allow for the verification of the authenticity of the information they contain. A second important finding is that the best method of ensuring ongoing authenticity of electronic records is external to the records themselves and involves a tight control on record-making and record keeping procedures and on the flow of metadata into the record's formal elements, rather than digital authentication measures, which have been proven to hamper long-term preservation of authentic records. These procedures and formal elements will be another deliverable of the research.

5. The Second Phase of InterPARES

In the course of the research, it has become apparent that the solutions identified for the long-term preservation of the administrative and legal records produced in large databases and document management systems are not sufficient for ensuring the continuing authenticity of records whose creation and form are discretionary, and which are generated by more complex systems. Therefore, the second phase of InterPARES, which will begin in January 2002, will focus on the reliable creation as well as on the authentic preservation of records in dynamic, interactive, performance, and experiential systems, including those produced in the course of creative and performing activities. This is especially urgent in light of several facts: first, some governments, including Canada and Italy, are expected to go entirely on-line in a very short term, and to carry out their transactions with the citizens through interactive websites; second, increasingly, the products of creative and performing activities originate in digital forms not controlled by any existing standard; third, the standards that are being developed in relation to the records generated using computer technology do not keep into account cultural diversity and pluralism; and fourth, both those who produce and those who use complex information systems appear to be uniquely concerned with the "here and now", showing a great disregard for the permanent preservation of a recorded authentic memory of our times. Administrative transparency, historical accountability, long term legal requirements and the protection of culture require that governments, universities and industry look beyond the present and consider the political, social and economical implications of entrusting all knowledge to digital systems destined to quick obsolescence before having in place strategies and standards for their continuing authentic preservation. The most important achievement of InterPARES has been to get experts from all sectors to work together in a sustained, intense, consistent and integrated way, irrespective of differences in culture, discipline and intent. But this is only the beginning of a necessary worldwide effort.

6. References

http://www.interpares.org

Council on Library and Information Resources, *Authenticity in a Digital Environment* (Washington, D.C.: CLIR, 2000)