

Authenticity Metadata and the IPAM: Progress toward the InterPARES Application Profile

Joseph T. Tennis
University of Washington
Information School, USA
jtennis@uw.edu

Corinne Rogers
School of Library, Archives,
and Information Studies,
UBC, Canada
cmrogers@interchange.ubc.ca

Abstract

The presumption of authenticity of records is predicated on capturing information about the records' identity and integrity over the course of the lifecycle. Extant metadata schemas do not capture the totality of this metadata required by International Research on Permanent Authentic Records in Electronic Systems (InterPARES). This paper presents research to date that addresses this need—the functional requirements, domain model, and sample properties from the IPAM (InterPARES Authenticity Metadata) schema.

Keywords: authenticity metadata; application profile; domain model; functional requirements; digital preservation metadata

1. Introduction

Metadata, the machine- and human-readable assertions about information resources, enables intellectual, physical, and technical control over information resources, and is primarily discussed in the context of information retrieval and preservation. However, there are other reasons to create metadata. Digital records are the lasting traces of actions and transactions, and their evidentiary capacity depends on their authenticity, reliability and accuracy, established at creation, and maintained and preserved over time and across technological change. While functional requirements for preservation metadata are transitioning from research to standardization and implementation (e.g., PREMIS), little work has connected archival theory with functional requirements for authenticity metadata (cf. MoReq, ISO, AGRKMS). However, there is a wealth of research into the requirements for authenticity and reliability of records at their creation, and throughout their life cycle, culminating in preservation (InterPARES, 2008). Many extant metadata schemas account for identity and integrity (the components of authenticity), but do not explicitly capture the attributes identified and required by InterPARES research. And although we may be able to repurpose some aspects of metadata schemas to capture identity and integrity metadata, it is incomplete (according to InterPARES requirements), and risks being confused with the original purpose of the repurposed metadata schema. This would violate the requirement for interoperability established within the context of InterPARES 2 and codified in InterPARES 3 (Tennis, 2010 and Tennis et al., 2010). This paper presents our research to develop a metadata schema that satisfies the requirements set out by InterPARES theory.

It is instructive to examine several schemas that have addressed metadata intended to support the presumption of authenticity to understand why further work is still needed. This section will provide a brief overview of four such schemas, mentioned above: PREMIS, MoReq, ISO 23081, and AGRKMS. PREMIS (Preservation Metadata: Implementation Strategies) is a working group sponsored by OCLC (Online Computer Library Center) and RLG (Research Libraries Group) whose goal is “to define implementable, core preservation metadata, with guidelines/recommendations for management and use” (PREMIS, n.d.). The PREMIS Data Dictionary defines preservation metadata as the information required by a repository to support

the process of preservation, specifically “maintaining viability, renderability, understandability, authenticity, and identity *in a preservation context*” (PREMIS 2011, 3) (authors’ italics). Demonstration of authenticity is linked in PREMIS with fixity, that is verification that a file has not changed since an earlier point in time, and integrity, verified through format identification and validation. Authenticity may be demonstrated through the maintenance of provenance information, the preservation of versions that are identical to the content originally submitted, and through digital signatures (PREMIS 2011, 209-210).

MoReq (Model Requirements for the Management of Electronic Records), first published in 2001, updated and rereleased as MoReq2 in 2008, and again as MoReq10 in 2011, is a model specification for procuring and for developing electronic records management systems (ERMS). At its core is the requirement that records stored in a MoReq-compliant ERMS are authoritative records as defined by ISO 15489, that is, records that have the characteristics of authenticity, reliability, integrity, and usability (MoReq2 2008, 21). The specific elements required to demonstrate authenticity, however, are not articulated.

The last two standards mentioned above are standards for records management metadata. ISO 23081, Information and Documentation – Records management processes – Metadata for records, outlines principles (Part 1:2006) and conceptual and implementation issues (Part 2:2009) for creating recordkeeping metadata, and guidance on conducting a self-assessment on records metadata (Part 3:2011) (see www.iso.org). ISO 23081 builds on ISO 15489, the standard for records management, adopting the definitions of records (information created, received, and maintained as evidence and information by an organization or person, in pursuance of legal obligations or in the transaction of business (ISO 15489, 3.16)) and authenticity (an authentic record is one that can be proven (a) to be what it purports to be, (b) to have been created or sent by the person purported to have created or sent it, and (c) to have been created or sent at the time purported (ISO 15489, 7.2.2)). In ISO 23081 we find high level guidance for recordkeeping metadata from the perspective of archival and records management.

The Australian Government Recordkeeping Metadata Standard version 2.0 (AGRkMS) is compliant with AS ISO 23081 and provides 26 metadata elements and an additional 44 sub-elements that may be used to describe five separate entities: record, agent, business, mandate, and relationship. This recordkeeping metadata thus describes the context, content, and structure of records and their management through time (AGRkMS 2008, 7). This standard also recognizes the description of authenticity as a requirement to ensure “reliable, meaningful and accessible records that satisfy business needs, evidential requirements and broader community expectations” (AGRkMS 2008, 9). Evidence of authenticity is explicitly identified in the elements ‘date range’ (p. 28), recordkeeping event name: migrates (p. 94), and document form name: digital signature (p. 105).

These and similar metadata schemas recognize the need for identity and integrity metadata to establish and assess the authenticity of digital resources. Their primary goal, however, is broader than a demonstration of authenticity alone – they focus on preservation, recordkeeping, recordkeeping systems, in which authenticity plays an important part, but is only one aspect of the schema. While this has been driven by practical concerns, there remains a gap in our understanding of the necessary and sufficient requirements for establishing and assessing the authenticity of digital records. That gap is what this current research is intended to fill.

In this paper we present our research to date: the functional requirements, domain models, and areas of description for authenticity metadata. We then compare these human- and machine-readable assertions with metadata from the domains of preservation and bibliographic control, employing a framework analysis approach (Tennis, 2006). Framework analysis allows us to compare structure, work practice, and purpose of different knowledge organization systems, in this case metadata application profiles. In this way it is clear to see how complex and complementary different regimes of metadata can be.

2. Background on Authenticity and Digital Records

We take as our starting point the concept of a record as understood in archival theory. “Record” is defined by the International Research on Permanent Authentic Records in Electronic Systems (InterPARES), as “a document made or received in the course of a practical activity as an instrument or a by-product of such activity, and set aside for action or reference” (InterPARES). Authenticity is defined as being “of undisputed origin, genuine; reliable or trustworthy” (Canadian Oxford Dictionary, 2004). An authentic record, therefore, is one “that is what it purports to be and is free from tampering or corruption” (InterPARES, 2005). So defined, a record has distinct attributes of identity and integrity, the expression of which must be inextricably linked to it through metadata in order to guarantee the presumption of authenticity at the moment of creation, and assure its reliability and accuracy as it is used, maintained, and ultimately, preserved.

InterPARES has developed knowledge essential to the long-term preservation of authentic records created and/or maintained in digital form, and provides the basis for standards, policies, strategies and plans of action capable of ensuring the longevity of such material and the ability of its users to trust its authenticity. The intellectual framework of InterPARES is grounded in archival science and diplomatics, international in scope, and supported by an interdisciplinary process that has included a wide range of academic and professional fields, from sciences and the arts, to computer engineering and law.

InterPARES has provided a framework for assessing the authenticity of digital records. Carried out in three phases, InterPARES researched the requirements for the preservation of authentic digital records in static, dynamic, and interactive systems. The research identified the following necessary characteristics of digital records, which, captured in metadata and archival description, form the basis for establishing and assessing their authenticity:

- fixed form
- stable content
- explicit linkages to other records inside or outside the digital system through classification code or other unique identifier
- an identifiable context of creation
- the involvement of five identifiable persons: author, addressee, writer, originator, creator
- an action in which the record participates or which the record supports procedurally or as part of the decision making process (Duranti, 2009b; Duranti & Thibodeau, 2006; MacNeil, 2000b).

Fixed form and stable content are self-evident concepts in paper records, but problematic to define and establish in the digital environment.

InterPARES 1 (1999-2001) was organized around four domains of inquiry for inactive electronic records: conceptual requirements for preserving authentic electronic records and the identification of elements necessary to maintain their authenticity over time; appraisal of authentic electronic records; procedures and resources necessary for long-term preservation of electronic records; and a framework for policies, strategies and standards for long-term preservation of authentic electronic records. Research was conducted from the point of view of the preserver and the life-cycle model of records.

In addition to dealing with issues of authenticity, InterPARES 2 (2002-2007) delved into the issues of reliability and accuracy during the entire lifecycle of records, from creation to permanent preservation. It focused on records produced in complex digital environments in the course of artistic, scientific and governmental activities. InterPARES 3 (2007-2012) built upon the findings of InterPARES 1 and 2, as well as of other digital preservation projects worldwide to

put theory into practice, applying the results of the previous two phases through case studies with small and medium-sized organizations.

The objective of InterPARES 2 was to develop a theoretical understanding of the records generated by experiential, interactive, and dynamic systems, of their process of creation, and of their present and potential use in the artistic, scientific, and governmental sectors (Duranti and Preston 2008, p.225). One result of the research was a function model, called the Chain of Preservation (COP) model, which depicts the activities, inputs, and outputs that are needed to create, manage, and preserve authentic and reliable digital records. A function model is a structured representation of the functions, activities, and processes within the modeled system (Duranti and Preston 2008, p.227). The COP model represents the life-cycle management of digital records, founded on the concepts, methods, and practices of archival science, and grounded in the context of the applicable juridical system—that is, the laws, regulations, and norms that govern or affect record processes (Duranti and Preston 2008, p. 230).

The determination and assessment of authenticity depends on the circumstances of record creation, and framework of subsequent preservation. According to InterPARES, to assess the authenticity of a digital object, one must be able to establish its identity and demonstrate its integrity. The identity of a digital object is established by the attributes of the object that uniquely distinguish it from other objects, while integrity refers to its wholeness and soundness, that is, the degree to which it is complete and uncorrupted. InterPARES recognized and articulated the difference between the form in which a document is viewed by a person reading it, and the form in which it is stored in the electronic system. The layers of abstraction introduced by the technology between the physical and logical record have implications for the assessment of authenticity. This significant difference between paper and digital records is at the root of the challenges of integrating digital records with the legal system's conception of documentary evidence.

Distinct from other disciplines where metadata is employed therefore, in the context of the InterPARES project, metadata are *assertions about documents, records and groups of records that we make for the sake of authenticity, preservation and retrieval*. Building on the theoretical work of InterPARES we have drafted a set of functional requirements, domain models, and areas of description for authenticity metadata. The attributes and linkages that we require for the presumption of authenticity are articulated in the *Benchmark Requirements Supporting the Presumption of Authenticity of Electronic Records* and *Baseline Requirements Supporting the Production of Authentic Copies of Electronic Records*, and modeled in the Chain of Preservation model. They include the names of all persons concurring in the formation of the record, the name of the action or matter, all dates of creation and transmission, the expression of the relationship of the record to related records (the archival bond), indication of attachments (these are the identity metadata), names of offices handling the record including the office with primary responsibility for it, indication of types of annotations,, indication of all technical modifications (these are the integrity metadata), as well as indication of rights and access privileges, and indications of various controls, policies, and procedures (InterPARES, 2008).

3. Methodology

This is design research. As design research, the methodology followed a course of knowledge acquisition: (1) establish the desired outcome for metadata schema; (2) follow the DCMI Singapore Framework; (3) develop functional requirements and model those through entity relationship diagrams in order to then construct an IP3 application profile; and (4) vet this with stakeholders, researchers, and the DCMI community.

The research was undertaken through the lens of archival theory in general, and the findings of InterPARES 1 and 2 specifically, applied in relation to the Guidelines for Application Profiles (Coyle & Baker, 2009) and the Singapore Framework (DCMI, 2008), both issued by the DCMI.

We were guided by literature on the creation of application profiles (Heery and Patel, 2000), and examples of application profiles (Collections and Scholarly Works Application Profiles).

4. Functional Requirements

In accordance with best practice, we established functional requirements for our metadata application profile (InterPARES authenticity metadata, or IPAM). Six functional requirements encompassed the scope of IPAM: (1) presumption of authenticity, (2) interoperability between systems and across time, (3) parsimony, (4) adequacy for archival description, (5) retrieval, and (6) meaningful display.

We can make clear what we want from our metadata by translating this set of functional requirements into a sentence. It reads, *these metadata should be necessary and sufficient to support the presumption of authenticity of records, interoperate between systems and across time, be adequate for archival description, and be useful for both retrieval and meaningful display of records.*

In order to implement the first two functional requirements, we adopted the expressions of record attributes and linkages identified in the InterPARES Benchmark Requirements Supporting the Presumption of Authenticity of Electronic Records. These establish the core information for the record's identity, and the foundation for demonstrating its integrity, including conditions of access, protection from loss, corruption and obsolescence, form and means of authentication. The last three functional requirements are realized in the application of Baseline Requirements Supporting the Production of Authentic Copies of Electronic Records. These are the minimum elements necessary to enable the production of authentic copies of the records, expressed through expression of controls, documentation, and contextual information. While not all creators or preservers of electronic records may require all elements, taken as a whole they represent the necessary and sufficient – in other words, the parsimonious set of metadata elements that will satisfy the most rigorous requirements for authenticity and international standards for archival description.

5. Domain Model

The domain model, in the context of metadata application profiles, describes in both words and in entity-relationship diagrams, what is to be described and how those things to be described relate to one another. In our case, we have constructed three levels of the domain. The middle layer is the *record*. The layer above the record is the *aggregation of records* and the layer below the record is *record attributes*. See Figure 1.

Diagram 1.0
Overview

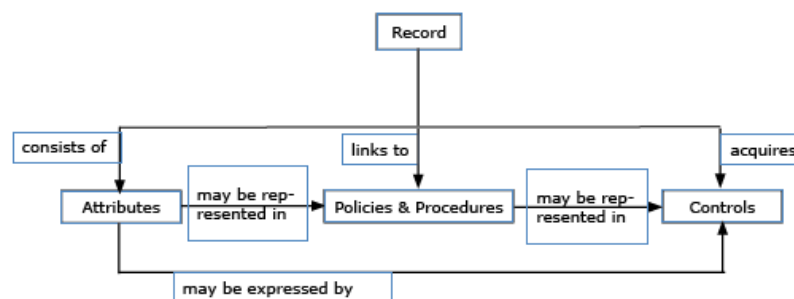


FIG. 1. The record in the domain model

The concept of “record” in the context of this research is based on the fundamental assumption of diplomatics, adopted by InterPARES, that it is possible to conceive of one typical, ideal documentary form that contains all elements of a record. This includes elements of documentary form, annotations, context, and medium. These are identified in the attributes, policies and procedures, and controls modeled in Figure 1 and Figure 2.

The record *consists of* attributes, *links to*, or *is linked to* policies and procedures, and *acquires* controls as it moves from document to individual record to part of an aggregation of records. Attributes *comprise* both intellectual and digital components. These are tightly related attributes linked to both the identity and integrity of the record in a digital system. And both statements about the identity and integrity of a record and its carrier (respectively) are required to assert, on behalf of the preserver, whether the record can be presumed to be authentic. Thus, any metadata that takes as its purpose the presumption of authenticity must document the identity and integrity of the record by recording the intellectual and digital components. The intellectual and digital components document the record’s identity, whereas the policies and procedures and controls primarily (though not exclusively) document the integrity of the record. See Figure 2.

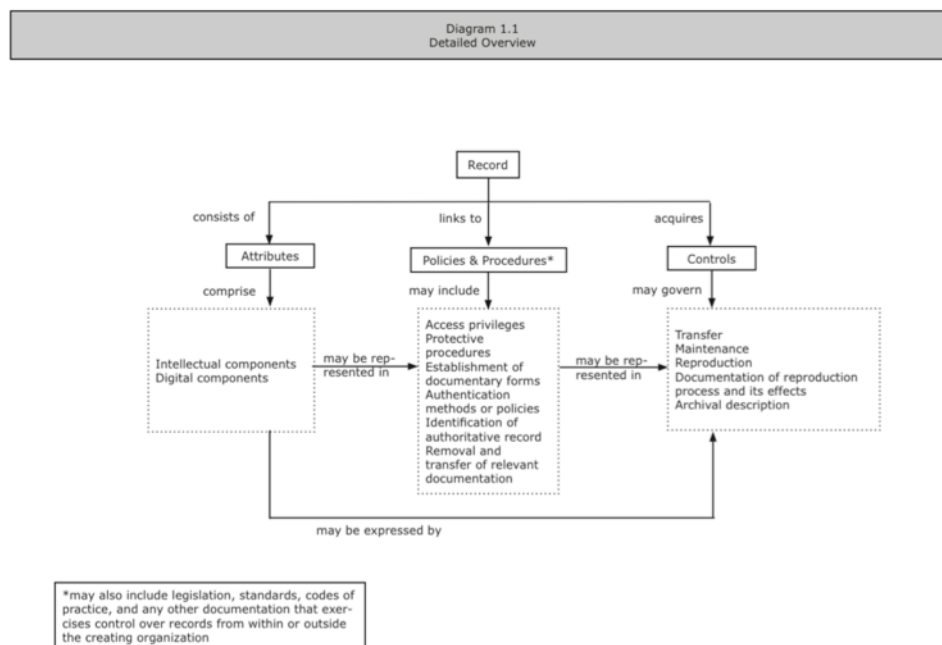


FIG. 2. The attributes, policies and procedures, and controls of a record

6. Development of Metadata Elements

Once the functional requirements were established and the domain models created, we developed the specific metadata elements for the application profile. These elements corresponded to the metadata elements identified from the COP model (Preston, 2009).

Metadata elements were coded by function and then numbered according to their appearance in the COP model. Functions were identified according to principles of archival and diplomatic theory as expressed by previous InterPARES research:

- AT – attachments: Signals those items attached to the record – indication of attachments is necessary for the integrity of the record.
- AU – authentication: Those elements that indicate the identity of the persons involved in the creation of the record.

- B – archival bond: Those elements that illuminate the connection of the record to other records to which it relates, and its context, whether it is preserved or destroyed.
- D – date: Points in time in the life cycle of the record(s) that need to be documented.
- DO – external documentation: Links to information that governs preservation, transfer, and access to the record(s) over time.
- F – form: The rules of representation that determine the appearance of an entity and convey its meaning.
- H – handling: Representation of the office or officer formally competent and/or responsible for carrying out the action to which the record(s) relates or for the matter to which the record(s) pertains.
- L – location: Indications of where the record(s) are stored, backed up, duplicated.
- P – persons: Identification of individuals or legally defined entities who are the subject of rights and duties and are recognized by the juridical system as capable of or having the potential for acting legally with regard to the record(s)
- R – rights and access: restrictions or privileges that apply to the record(s).
- S – subject: The action or matter to which the record(s) pertain.
- T – technology: The carrier(s) of the form and content of the record.

Once the metadata elements were enumerated to match the COP model and grouped according to function, we crosswalked our elements with MoReq 2010, Dublin Core, and PREMIS.

The next step was to interpret the metadata elements using the World Wide Web Consortium's (W3C) Resource Description Framework (RDF) for eventual publication and review in the wider community. The purpose of encoding metadata in this form is so that it is usable by information systems, but also human readable.

7. RDF Binding of IPAM

After we finalized the spreadsheet we then bound the statements in RDF using Google Refine. We then exported the file. Here is a sample of the RDF generated from this process.

```
<rdf:Description rdf:about="http://ipam.info#A2.2.2.0">
  <rdf:type rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Property"/>
  <rdfs:isDefinedBy
    rdf:resource="http://www.interpares.org/ip2/ip2_model_display.cfm?model=cop"/>
  <rdfs:label xml:lang="en">chronDate00</rdfs:label>
  <dcterms:description xml:lang="en">The date of document creation</dcterms:description>
  <rdfs:domain rdf:resource="http://ipam.info#Record"/>
  <dcterms:identifier rdf:resource="http://ipam.info#D00"/>
</rdf:Description>
```

Here we are describing a resource, which is an attribute, state, point in time, or agent or documentation associated with a record or group of records. In this case we are talking about a date of creation. This date of document creation is positioned in the chain of preservation using its resource name A2.2.2.0, which is defined by the model at www.interpares.org. We provide a label and a description, and identify the domain of the property. In this case it pertains to a record. Other ranges include transfer actions (moving records from one system to another), and various reports on the feasibility of preservation, etc.

8. Future Work

Future work will involve building out this metadata in systems, and testing the application of authenticity metadata in digital archives. One promising venue is the work being done by Artefactual Systems on their open source Archivematica software (Artefactual, 2012).

Maintaining the presumption of authenticity of digital records is a pressing concern for society. Trusting our digital repositories requires that we invest in efforts to document the means of

creation, transactions, transformations, and states of digital records. The full contextualization of records that allows us to trust their authenticity and reliability is achieved through the capture and maintenance of complete archival metadata for authenticity. Using theory derived from the literature and collaborative research projects, as well as methods outlined by the DCMI, we have taken one small step toward aiding that effort.

References

- Artefactual Systems (2012). <http://www.artefactual.com/>
- National Archives of Australia (2008). "Australian Government Recordkeeping Metadata Standard version 2.0." Available at www.naa.gov.au.
- Baker, T., M. Dekkers, R. Heery, M. Patel and G. Salokhe (2001), "What terms does your metadata use? Application profiles as machine understandable narratives," *Journal of digital information* 2(2).
- Canadian Oxford Dictionary, 2nd ed. (2004). (Don Mills, Ontario : Oxford University Press.)
- Coyle, K. & T. Baker. (2009). Guidelines for Dublin Core Application Profiles. Available at <http://dublincore.org/documents/profile-guidelines/index.shtml>
- InterPARES 1: Authenticity Task Force (2005), "Appendix 2: Requirements for Assessing and Maintaining the Authenticity of Electronic Records" in *The Long-term Preservation of Authentic Electronic Records: Findings of the InterPARES Project*, Luciana Duranti, ed. (San Miniato, Italy: Archilab, 2005), 204–219. Online reprint available at http://www.inter pares.org/book/inter pares_book_k_app02.pdf.
- Dublin Core Metadata Initiative (DCMI) (2008). "The Singapore Framework for Dublin Core Metadata Application Profiles." Available at <http://dublincore.org/documents/singapore-framework/>. Last accessed March 18, 2012.
- Duranti, L. and Preston, R., eds. (2008). *International Research on Permanent Authentic Records in Electronic Systems (InterPARES) 2: Experiential, Interactive, and Dynamic Records*. Padova: Associazione Nazionale Archivistica Italiana.
- Duranti, L. and Thibodeau, K. (2006). "The Concept of Record in Interactive, Experiential, and Dynamic Environments: the View of InterPARES." *Archival Science* 6: 13-68.
- European Commission (2008). Model Requirements for the Management of Electronic Records – Update and Extension 2008. Luxembourg: Office for Official Publications of the European Communities. Available at <http://www.moreq2.eu/moreq2>.
- Heery, R. and Manjula P. (2000), "Application profiles: mixing and matching metadata schemas," *Ariadne* 25 (24 September 2000). Available at <http://www.ariadne.ac.uk/issue25/app-profiles/intro.html>.
- International Organization for Standardization (2001). ISO 15489-1:2001 Information and Documentation – Records management.
- _____. ISO 23081-1:2006 Information and Documentation – Records management processes – Metadata for records. Part 1: Principles.
- International Research on Permanent Authentic Records in Electronic Systems (InterPARES) 2: Experiential, Interactive and Dynamic Records, Luciana Duranti and Randy Preston, eds. (Padova, Italy: Associazione Nazionale Archivistica Italiana, 2008). http://www.inter pares.org/display_file.cfm?doc=ip2_book_introduction.pdf.
- Nilsson, M., T. Baker and P. Johnston (2008), Singapore Framework for Dublin Core Application Profiles. Available at <http://dublincore.org/documents/singapore-framework/>.
- PREMIS (N.d.). "Information Sheet." Available at www.loc.gov/standards/premis/information-sheet.pdf.
- Preston, Randy (2009). "InterPARES 2 Chain of Preservation (COP) Model Metadata (Draft)." http://www.inter pares.org/rws/display_file.cfm?doc=IP2-cop-model_metadata_v1.0.doc.
- Scholarly Works Application Profile (SWAP). Available at http://www.ukoln.ac.uk/repositories/digirep/index/Eprints_Application_Profile.
- Tennis, J. T., C. Rogers, S. Xie and R. Preston (2010), "Report on Metadata Developments (1): Interoperability Across Systems, Time, and Conceptions," InterPARES 3 Project, Fourth International Symposium, 17 September 2010. Oslo, Norway. Available at http://www.inter pares.org/display_file.cfm?doc=IP3_isym04_presentation_4-2_canada.pdf.
- Tennis, J. T. (2010), "Metadata: Interoperability Across Systems, Time, and Conceptions," InterPARES 3 Project, Third International Symposium, 29 May 2010. Vancouver, British Columbia, Canada. Available at http://www.inter pares.org/display_file.cfm?doc=ip3_isym03_presentation_04-04--tennis.pdf.