Autorità per l'informatica nella Pubblica Amministrazione

La conservazione dei documenti informatici - Aspetti organizzativi e tecnici

Seminario di studio

Roma 30 ottobre 2000

Programma

I SESSIONE: LA FUNZIONE CONSERVATIVA. PROBLEMI, STRATEGIE, ESPERIENZE,

9:00 **Apertura del seminario**

Salvatore Italia Direttore Ufficio Centrale per i beni archivistici

Guido Mario Rey –Presidente dell'Autorità

9:15 **Relazione introduttiva**

Carlo Batini (Autorità per l'Informatica nella Pubblica Amministrazione)

9:35 **La gestione e la conservazione dei documenti informatici nelle pubbliche amministrazioni italiane (The management and the preservation of electronic records in the Italian public administrations)**

Antonio Massari (Autorità per l'Informatica),

10:10 **Information Management Architecture for Persistent Object Preservation**

Reagan W. Moore (University of California)

10:40 **Intervallo**

11:00 **The strategies for the preservation. Projects and perspectives in the US Federal Government**

Ken Thibodeau (National Archives USA)

11:30 **The Swedish programs for the electronic records preservation and access**

Kristiansson Gorlan (Riksarkivet, Sweden)

12:00 **Responding to the challenges and opportunities of ITC: the new record manager**

Seamus Ross (University of Glasgow)

12:30 **Discussione**

13:00 **Intervallo per il pranzo**

II SESSIONE. I METODI PER LA CONSERVAZIONE, IL RUOLO DEI METADATI, LE POTENZIALITÀ DI XML

14:00 **Relazione introduttiva**

Maria Guercio (Università di Urbino)

14:30 **A logical model for the electronic records authentication**

Bill Underwood (Georgia Tech Institute)

15:00 **Metadata: Archival Concept or IT Domain?**

Peter Horsman (IT Committee, International Council on Archives, president)

15:30 **The potential of markup languages to support descriptive access to electronic records: the EAD standard**

Anne Swetland Gilliland (University of Los Angeles)

16:00 **XML standards for business-to-business and business to government communication**

Zachary Coffin (eXtensible Business Reporting Language)

16:30 **XML, uno standard per gli archivi informatici (XML, a standard for the electronic records)**

Daniele Tatti (Autorità per l'informatica)

Indice

Premessa

Il seminario, organizzato dall'Autorità e dall'Ufficio Centrale per i Beni Archivistici del Ministero per i Beni e le Attività Culturali, ha la finalità di affrontare la questione cruciale della conservazione a lungo termine dei documenti informatici, da un lato sensibilizzando le amministrazioni pubbliche che hanno già avviato o programmato processi di automazione dei sistemi documentari, sia di avviare - anche attraverso un confronto internazionale - l'analisi e la verifica delle modalità e degli strumenti che consentono il mantenimento nel tempo dell'integrità e dell'autenticità delle memorie digitali nonché la loro accessibilità. Il tema ha acquistato attualità e rilevanza crescenti in relazione alla larga diffusione di tecnologie dell'informazione e della comunicazione per la gestione dei flussi amministrativi e documentari. Richiede , tuttavia, la soluzione di numerosi problemi di natura organizzativa e tecnica che gruppi internazionali di ricerca e istituzioni nazionali di tutto il mondo hanno cominciato a studiare da qualche anno con impegno crescente soprattutto allo scopo di individuare soluzioni scalabili e di basso costo che rendano possibile anche alle organizzazioni di piccole e medie dimensioni avviare interventi generali di automazione dei processi amministrativi e delle attività di trasmissione e tenuta dei documenti.

In particolare, **nella prima sessione**, il seminario si propone di offrire alle amministrazioni pubbliche italiane - soprattutto ai responsabili dei sistemi informativi e ai responsabili dei sistemi documentari - un panorama della specifica situazione nazionale e dei progetti e delle esperienze operative più interessanti condotte in altri Paesi (Stati Uniti e Svezia) in stretta collaborazione con più ampi gruppi di ricerca internazionali, insieme a un'analisi delle esigenze di formazione e riqualificazione dei profili professionali destinati ad assumere il nuovo difficile ruolo di conservatore delle memorie digitali.

La seconda sessione è destinata ad esaminare le questioni più strettamente tecniche della conservazione, in particolare i diversi metodi oggi disponibili per affrontare l'obsolescenza tecnologica e, in special modo, le potenzialità dei linguaggi di marcatura (SGML, XML) per il trattamento e la gestione dei metadati che garantiscono la tenuta a lungo termine del patrimonio documentario e lo sviluppo di standard specifici (Encoded Archival Description).

The potential of markup languages to support descriptive access to electronic records: The EAD standard

Anne J. Gilliland-Swetland

Abstract:

This paper will review the potential of Encoded Archival Description (EAD), rcently adopted as an American descriptive standard, to provide online descriptive access to electronic records. The paper will begin by reviewing the current state of electronic records description and the complex relationships between metadata that are part of the record and metadata that are about the record. It will then describe the status and scope of EAD, how it relates to other descriptive initiatives that are applying markup languages, and the potential of EAD to serve as a metadata infrastructure for online archival information systems. The paper will conclude with a discussion of the extent to which EAD can currently accommodate, or could be extended to accommodate, description and online delivery of electronic records.

Introduction

There has been a considerable amount of political and professional rhetoric, stemming from unprecedented developments over the past decade in technologies supporting the World Wide Web, about developing online access to unpublished information resources—including archival holdings. The rhetoric has resulted in the establishment of research and development agendas by major government funding agencies, private foundations, industry, and professional institutions and associations[4]. A number of major initiatives have resulted from the availability of this funding. As they relate to archival concerns, these initiatives can be grouped into three primary domains of activity:

- the development of archival standards that support online access to archival descriptions (Encoded Archival Description being the most prominent recent example);

- the development of archival information systems such as American Memory at the Library of Congress and the Online Archive of California (to cite two American examples) that provide not only online descriptions but also digitized copies of selected archival holdings; and

- research projects addressing the archival management of records that are "born digital," that is, of electronic records (for example, the Recordkeeping Functional Requirements Project at the University of Pittsburgh and the International Project on Permanent Authentic Records in Electronic Systems (InterPARES))[5].

While there has been considerable dialog and overlap between archivists involved with the first two of these areas, until recently archivists grappling with the challenges of creating and preserving electronic records have not been integrally engaged in broader initiatives to standardize and enhance description for online access, nor to provide online access to electronic records through archival information systems or digital libraries. The major exception to this has been the Recordkeeping Metadata Schema (RKMS)

[4] For example, the National Science Foundation, the National Endowment for the Humanities, and the National Historical Publications and Records Commission in the United States, and the Fifth Framework and the Joint Information Systems Committee in Europe; national archives and libraries in many countries; and descriptive standards groups within professional associations.

[5] Gilliland-Swetland, Anne J. and Philip Eppard. "Preserving the Authenticity of Contingent Digital Objects: The InterPARES Project" *D-Lib Magazine* 6 no.7 (2000). Available at: http://www.dlib.org/dlib/july00/eppard/07eppard.html (16 October, 2000);
InterPARES Website available at http://www.interpares.org (16 October, 2000).

developed in Australia. The focus for RKMS is the record, regardless of its format, and how it can be reconstructed and retain its meaning across time and user domains. RKMS provides:

- A standardized set of structured or semi-structured recordkeeping metadata elements
- A framework for developing recordkeeping metadata sets in different contexts
- A framework for mapping recordkeeping metadata sets to establish equivalences and correspondences that can provide the basis for semi-automated translation between metadata sets[6]

In North America, electronic records management evolved to some extent as an area apart from the mainstream of the archival profession. Its immediate concerns have been on creating, identifying, and accessioning electronic records. In the 1970s and 1980s, electronic records, or "machine-readable records" as they were initially termed, tended to be managed as software-independent datafiles. More recently, as electronic records have taken on more complex functionality, there has been an increased awareness of the need to preserve their value as legal and organizational evidence. As a result, archivists are now engaged with researchers from computer science, digital library development, and preservation in several projects to identify how to preserve authentic electronic records with their functionality intact. One of the most prominent of such projects is that of the National Archives and Records Administration and the San Diego Supercomputer Center to employ XML in the development of persistent archives. This concern for evidence requires a more detailed understanding of what are the characteristics of an authentic record in and over time, as well as close analysis of the intellectual rationales behind archival description in terms of how it contributes to ensuring and demonstrating the authenticity of preserved records.

Indeed, there is a growing convergence of different areas within the archival profession, as well as of other professional and disciplinary domains relating to description. This convergence arises largely out of the development of new metadata schema and standards and technological capabilities that provide structures and crosswalks[7] for formalizing and bridging diverse data types (such as image or geospatial data), metadata semantics, and professional practices[8].

Archives play a key and often overlooked role in establishing and demonstrating the authenticity of any record, regardless of its form, through archival description. In contrast to the key purposes of bibliographic description which are to manage a physical information object as well as to facilitate its intellectual retrieval and use, archival description must address that object not only as information, but as evidence. As a result, archival description must not only describe the content of a fonds or record group, it must also describe the circumstances of its creation, its chain of custody, its relationships to other records generated by the same activity, and the impact upon the aggregation of records of any processing or preservation activity in ways that are and remain meaningful to different kinds of users over time. Archival description, therefore, has three primary roles. Firstly, it serves as a tool that meets the needs of the archival materials being described by authenticating and documenting them. Secondly, it is a collections management tool for use by the archivists. Thirdly, it is an information discovery and retrieval tool for making the evidence and information contained in archival collections available and comprehensible by archivists and users alike.

---

[6] See McKemmish, Sue, Glenda Acland, Nigel Ward, and Barbara Reed. "Describing Records in Context in the Continuum: The Australian Recordkeeping Metadata Schema." *Archivaria* 48 (1999): 3-42.

[7] A crosswalk is a chart or table that represents the mapping of fields or data elements in one metadata standard to fields or data elements in other standards that have the same function or meaning. Crosswalks support the ability to search transparently heterogeneous databases as a single database (semantic interoperability) and to convert data from one metadata standard to another.

[8] See Gilliland-Swetland, Anne J. *Enduring Paradigm, New Opportunities: The Value of the Archival Perspective in the Digital Environment* (Washington, D.C.: Council on Library and Information Resources, 2000).

Describing Electronic Records

Ironically, in a world of increasing online access to primary information resources, many of which first require digitization, electronic records are proving to be among the most intransigent in terms of providing even basic descriptive access. This intransigence reflects inherent technical problems with the diverse formats in which electronic records are created and may need to be maintained. Equally, it reflects how the enormous volume of electronic records requiring processing by a comparatively small staff together with data archiving practices originally adopted from the social sciences data archives community have led to idiosyncratic summary archival descriptions and an over-dependence upon the metadata generated by the creator of the records. Description of electronic records often consists of high level summaries of data, reports on quality and accuracy of data, scanned or PDF versions of codebooks and data dictionaries, and customized subject indexes and data extracts.

While the current state of description for electronic records is certainly understandable, it is, nevertheless, deficient in several respects:

- There has been insufficient analysis of what is the actual nature of electronic records. In particular, there needs to be more examination of the relationship between data content and the metadata that provide and document its context and structure, and of the various ways in which aspects of data and metadata in complex systems such as databases might come together to form the intellectual construct that is a record. Often one of the most difficult aspects of working with electronic records is to be able to identify and then describe, in the absence of a tangible document, the parameters of that intellectual construct.

- Metadata generated by records creators has been viewed as sufficient substitute for archival description. For example, in 1993, Margaret Hedstrom proposed that management of metadata provide an alternative strategy to current descriptive practices in order to support the "need to identify, gain access, understand the meaning, interpret the content, determine authenticity, and manage electronic records to ensure continuing access[9]." Subsequently, several projects have resulted in metadata specifications for electronic records, most notably the Pittsburgh Project and related implementation projects such as the Indiana University Electronic Records Project. With the exception of the Australian RKMS project, there has been almost no discussion of the value-added role that archival description should play in terms of ensuring and documenting authenticity, and making the records meaningful to users across time and domains.[10]

- There has been little emphasis on establishing the documentary relationships between electronic records and paper records created by the same activity. Lack of standardization and use of non-archival of descriptive practices has made it difficult to integrate descriptions of electronic records with standardized descriptive metadata created by archivists and other information, industry, and research communities. For example, in the mid-1980s, when archivists looked to the use of MARC formats to the MARC Machine-Readable Data Format (MRDF) rather than the MARC Archives and Manuscripts Control Format (AMC) that was developed for the collective description of archival and manuscript

[9] Hedstrom, Margaret. "Descriptive Practices for Electronic Records: Deciding What is Essential and Imagining What is Possible," *Archivaria* 36 (Autumn 1996): 53.

[10] Bearman, D. and Sochats, K. *Metadata Requirements For Evidence*. 1996, Available: http://www.lis.pitt.edu/~nhprc/BACartic.html (October 17, 2000); Bantin, Philip C. "Developing a Strategy for Managing Electronic Records: The Findings of the Indiana University Electronic Records Project," *American Archivist* 61 (1998): 328-64. Bantin, Philip C. "The Indiana University Electronic Records Project Revisited," *American Archivist* 62 (1999)153-163; and McKemmish, Sue, Glenda Acland, and Barbara Reed. "Towards a Framework for Standardising Recordkeeping Metadata: The Australian Recordkeeping Metadata Schema," *Records Management Journal* 9 (1999): 177-202.

materials. In effect, such an approach treated electronic records as a special format with distinct descriptive needs, rather than as components of wider archival aggregations.

- Because management of electronic records has generally been viewed by the rest of the archival profession as an area that requires distinct technical expertise, developments in archival description such as EAD have progressed without being strongly informed by the descriptive needs of electronic records.

It is useful at this point to define more closely what is meant by metadata, since the term is understood differently by different communities. Metadata refers to a range of structured or semi-structured data about data that are critical to the development of effective, authoritative, interoperable, scaleable, and preservable information and record-keeping systems. Until the mid-1990s, metadata was a term most prevalently used by communities involved with the management and interoperability of geospatial data, and with data management and systems design and maintenance in general. For these communities, metadata referred to a suite of industry or disciplinary standards as well as additional internal and external documentation and other data necessary for the identification, representation, inter-operability, technical management, performance, and use of data contained in an information system. For archivists, metadata refers to the value-added information, such as EAD, that they create to identify, authenticate, arrange, describe, preserve and otherwise enhance access to their holdings.

In contemplating the role of metadata in the description of electronic records, several questions come to mind:

- Which metadata are part of the record, which are about the record, and which are neither but are required to preserve or reconstruct the technological context of the record? And of all these types of metadata, which must be captured as part of archival description?

- How can the trustworthiness of these metadata be determined in terms of quality and completeness in and over time?

- Are their descriptive needs of electronic records that might be different from those of other types of records? If so, what are they and how should they best be addressed?

- Can the metadata generated by the creator of the electronic record somehow be automatically translated or mapped into a standardized description for archival records?

- Can the structure and documentary contexts of electronic records be automatically analyzed to generate specific components of a standardized description for electronic records?

- Which kinds of contextual documentation do electronic records require in order to be understood and can a metadata infrastructure facilitate links to that documentation online?

- How can the links between records and metadata retain their referential integrity over time and in the face of systems obsolescence, data migration, and evolution of metadata schema?

- What do users need in order to be able to identify relevant electronic records online? What do users need to be able to use electronic records disseminated online?

Encoded Archival Description

In the face of such questions, therefore, how might Encoded Archival Description and other markup initiatives enhance current electronic records description? Simply defined, EAD is a Document Type Definition (DTD) developed using Standard Generalized Markup Language (SGML) that makes it possible to develop predictably structured archival description that can be disseminated on the World Wide Web.

That description is most commonly an archival finding aid, but the DTD is flexible enough to accommodate various other types of archival descriptive tools.

However, the power of EAD is that it can be much more than a structure through which to create a digital representation of a two-dimensional finding aid. The hierarchical nature of EAD, its explicit delineation of each data element, and its adherence to standardized metadata conventions and protocols provide it with the potential to function as a multi-dimensional metadata infrastructure that can interface with other metadata schema, but that can provide maximum flexibility in describing a diversity of record types. With such an infrastructure, archivists and software developers have the capabilities and incentives to design a range of archival information systems that fundamentally re-conceptualize how access to archival holdings is provided. These archival information systems would not only contain the kinds of archival description found today in finding aids, but also digitized versions of archival materials, full-text of ancillary materials, extensive linkages to other online archival and bibliographic information systems, and actual electronic records and the necessary technical documentation to use them[11]."

In such information systems, however, EAD would not be the only metadata schema invoked, and one of the powerful aspects of EAD is its ability to interface or interoperate with other metadata schema and SGML-based implementations. EAD is fully XML-compliant, meaning not only that EAD-encoded descriptions can be searched and manipulated over the Web as the Web increasingly supports XML, but also that electronic records technical documentation, such as database models, workflow rules, and technical drawings can integrated with the archival descriptions in ways not previously possible in a more manual environment. Similarly, EAD can interface with descriptive metadata created in MARC because of metadata mapping between the two standards. With the recent release of XMLMARC software, this mapping will become only easier. EAD also shares header data elements with the Text Encoding and Interchange (TEI) DTD. TEI is a DTD that facilitates the development of digital versions of scholarly texts.

Using EAD to Describe Electronic Records

EAD is currently in its first full release (Version 1.0). It is fully expected that the DTD will be dynamic and will continue to be extended to accommodate new technological capabilities and metadata schema, as well refined based one evaluative feedback from archivists and users. In its current form, what then does EAD have to offer electronic records description, given that the needs of electronic records have yet to be integrally addressed by the DTD?

EAD, while it is a data structure and not a data content standard, works to standardize idiosyncratic descriptive practices. Electronic records descriptive practices are some of the most idiosyncratic in the field because there is such diveristy of types of electronic records, and because electronic records description is rarely taught in archival education programs, and is primarily learned as institution-specific practices "on the job." Descriptive records tend to comprise examples of descriptions of datafiles, rather than complete descriptions, together with user guides of documentation packages[12]. Using EAD would also integrate electronic records management into the mainstream of archival activities, treating the records as records, rather than as instances of special formats. Moreover, through collective description, as well as elements such as <separatedmaterial> and <relatedmaterial>, all records created by the same activity will be treated as an intellectual whole, regardless of whether they are paper, electronic, or some other medium.

---

[11] Gilliland-Swetland, Anne J. "Popularizing the Finding Aid: Exploiting EAD to Enhance Online Browsing and Retrieval in Archival Information Systems by Diverse User Groups" *Journal of Internet Cataloging* 4 nos. 1/2 (2000) (in press); Gilliland-Swetland, Anne J. "Health Sciences Documentation and Networked Hypermedia: An Integrative Approach," *Archivaria* 41 (1995): 41-56.

[12] Dryden, Jean E. "Archival Description of Electronic Records: An Examination of Current Practices," *Archivaria* 40 (1995): 99-108.

Electronic records descriptions can be quite flat, consisting mostly of summary information, with arrangement of the contents of a datafile often being incidental. However, users may wish to have access at the level of individual records or even data elements. The hierarchy built into EAD has the potential to support this kind of granularity of access, although commercial software that is currently available has yet to address much of this potential. Technical documentation accompanying the electronic records can also be linked in electronic form to the EAD description through elements such as <archref>, <odd> (other descriptive data) and <add> (adjunct descriptive data). If this documentation is marked up using SGML, XML, or some other markup language, the possibility exists of additional reconciliation if the different metadata schema. The well-defined EAD structure also makes possible the use of cross-walks to interface with other common metadata schema that might be relevant to the records (for example, geospatial metadata).

All this is not to say that EAD is ideal as it stands for describing electronic records. Several limitations need to be addressed in the next version of EAD if it is truly to accommodate electronic records.

1) EAD is strongest with regard to the description of the records once they are held in the archives. It is weak in how it supports records management, appraisal, and accessioning processes. More explicit attention needs to be paid to how records retention schedules, appraisal reports, accessioning procedures, and data quality reports are captured and tracked, as well as the various agents associated with those processes.

2) There needs to be a more closely delineated data elements with which electronic record metadata can be described, rather than consigning such materials to “bucket” elements such as <odd> and <add>. These elements and their values should be based upon lists of common types of documentation that accompany electronic records when they are accessioned. The data elements also should have attributed that indicate the extent to which the accuracy of each piece of documentation has been verified.

3) Custodial history is integral to establishing the authenticity of records, and for electronic records it can be quite complex, especially if the archives takes over intellectual and not physical control of inactive records. The <custodhist> element needs to be expanded to address this issue, in particular, non-custodial arrangements for electronic records.

4) Preservation and meticulous documentation of preservation processes are integral not only for providing continued access to electronic records, but also for establishing and demonstrating the continued authenticity of those records (or of authentic copies of the records). Currently preservation information is bundled into a single data element <processinfo> (processing information), and as with <custodhist> this element needs to be expanded and further delineated to track preservation processes such as migration and emulation and any effects these might have upon the record.

5) Even with traditional records, many archivists find it difficult to make the necessary distinction between intellectual and physical levels of arrangement. Many electronic records can be arranged in multiple ways and, therefore, the concept of levels of arrangement may not be as relevant as possible arrangement schema. It needs to be possible through the <arrangement> element for users to identify the range of potential arrangements and data extracts in order to be able to specify the one which they would like to use when accessing electronic records online or ordering copies of them. This is a compelling reason to do more user-based research so that any extensions to EAD are more user-driven.

6) As with museum objects, additional aspects of physical description may need to be incorporated into the <physdesc> element to allow for highly technical description. Some of these elements might correspond to those that were included in MARC MRDF.

7) For EAD in general, there is a need for a companion content standard and a structure for developing authority files. Work on both of these aspects is currently underway. There is also a need to analyze the extent to which EAD should be extended to accommodate a range of archival descriptive

traditions and technical requirements for records in specific media, or whether a better approach would be to concentrate on mapping different types of metadata through processes such as metadata crosswalks and automatic reconciliation of diverse XML structures.

Conclusion

There is obviously much work to be done in the area of electronic records description, and EAD provides one important vehicle to do so. However, given the volume of electronic records already created and anticipated in future years, there must surely also be an increased emphasis on automating as many aspects of archival description as possible. This is where research and development such as that underway at the San Diego Supercomputer Center in partnership with the US National Archives and Records Administration is likely to make such a strong contribution. One final caveat, however—almost all developments in archival description to date, even that of EAD, have occurred without systematic analysis of user needs and capabilities. As archival description, and even the complete archival record becomes increasingly available online to the general public without any archival reference mediation, it is going to be critical that we spend time examining the usefulness and usability of the materials we are providing to our users. Otherwise we may find that we have created a web of metadata and records that is so complex that it will have become impenetrable to most users.