European Commission

INSAR
Information Summary on Archives

European
Archives
News

SUPPLEMENT VII

Proceedings
# of the DLM-Forum 2002

@ccess and preservation of electronic information: best practices and solutions

DLM-FORUM
Electronic Records

Barcelona, 6–8 May 2002

EUROPEAN
ARCHIVES NEWS
INSAR

- Secretariat-General of the
  European Commission

- Editorial Committee
  Supplement VII:
  Peter BERNINGER
  Frank BRADY
  Hans HOFMANN
  Jef SCHRAM

- For further information:
  European Commission
  Tel. (32-2) 29-50557
  (32-2) 29-56721
  Fax (32-2) 29-61095
  E-mail: archis@cec.eu.int
  Address: rue de la Loi 200
  Office: JECL 3/35
  B-1049 Brussels

Proceedings

# of the

DLM-Forum 2002

@ccess and preservation of electronic information: best practices and solutions

Barcelona, 6–8 May 2002

Cataloguing data can be found at the end of this publication.

Luxembourg: Office for Official Publications of the European Communities, 2002

ISBN 92-894-4415-0

Printed in Italy

PRINTED ON WHITE CHLORINE-FREE PAPER

Knowledge and action for digital preservation: Progress in the US Government

Kenneth Thibodeau

In 1998, the National Archives and Records Administration (NARA) of the United States launched a new initiative to tackle the daunting challenges posed by electronic records. Although NARA had over 30 years of experience in this area, it was clear that the functional capabilities and quantitative capacity it had developed in that time were insufficient to cope with the ever-increasing variety and complexity, and the exponentially growing quantities of electronic records being produced by the US Government. The situation elsewhere was not particularly encouraging. NARA had extensive knowledge of what had been accomplished and was being explored by other institutions around the world, gained through participation in activities such as the International Council on Archives, the Interpares project, and the development of the open archival information system (OAIS) reference model. The state of affairs in 1998 could easily be summarised:

- proven methods for preserving and providing sustained access to electronic records were limited to the simplest forms of digital objects;
- even in those areas, proven methods were incapable of being scaled to a level sufficient to cope with the expected growth of electronic records; and
- archival science had not responded to the challenge of electronic records sufficiently to provide a sound intellectual foundation for articulating archival policies, strategies, and standards for electronic records.

In this environment, it seemed prudent to pursue a strategy of divide and conquer, partitioning the challenge of electronic records into manageable segments. The problem of the increasing diversity of electronic records was due to the different formats or data types in which digital data is encoded. To cope with this problem, NARA articulated a 10-year plan addressing different classes of data types sequentially in a three-step process. First, the characteristics of each class would be analysed. Then, options for preserving those characteristics would be evaluated. Finally, the best option would be chosen as the basis for developing the capability for preserving records in that class.

Parallel to this, NARA initiated a special project to address the worst-case scenario, which we had already encountered several times; namely, where the archives receive important bodies of digital objects which had not been managed as records, or had been poorly managed. In several of these cases, NARA had received obsolete computer systems and backup media from organisations which had ceased to function. In such cases, there is a basic and difficult problem of culling the records from the mass of other files, such as operating systems, application software, tutorials, help files, and temporary system files. To attack this problem, NARA sponsored a research project in collaboration with the US Army Research Laboratory and Georgia Tech Research Institute. This project focused on the contents of over 550 PC hard drives from the White House during the Administration of the former President Bush. It has developed a pilot system — called the Presidential Electronic Records PilOt System (Perpos) — of automated tools for identifying the data type of each file in a system or stored on media. On this basis, user-created files can be distinguished from software and related files. The project is investigating the application of advanced technologies to characterise the records and identify significant contents.([1])

To improve the knowledge base for digital preservation, NARA continued its support for development of the OAIS standard, and became one of the principal supporters of the International Research on Preservation of Authentic Records in Electronic Systems (Interpares) project. The OAIS effort had the clear advantage of bringing experts from many, independent disciplines and from many countries together to articulate the generic requirements for a system capable of maintaining and delivering information over time. Furthermore, because it was intended as an international standard, it had the potential for influencing the development of information technology products suitable for implementing an OAIS.([2]) The Interpares project had the advantage of focusing the efforts of ex-

Kenneth Thibodeau

Kenneth Thibodeau is a widely recognised expert in the management and preservation of electronic records. With 26 years experience in the field, he has spoken at over 120 conferences worldwide and has published more than 20 articles and chapters, in two languages, on these topics. He is Director of the Electronic Records Archives Programme at the National Archives and Records Administration in Washington and chairs the Preservation Task Force of the Interpares Project. Among other accomplishments, he directed the US Department of Defence Task Force that developed the DoD Standard 5015.2 for records management applications and served as secretary of the Committee on Electronic Records of the International Council on Archives and principal editor for the Committee's Guide to managing electronic records from an archival perspective. A fellow of the Society of American Archivists, he holds a Ph.D. in the history and sociology of science.

([1]) http://perpos.gtri.gatech.edu/

perts from a variety of disciplines and countries on the specific problem of guaranteeing the authenticity of electronic records over time. Since 1998, this project has articulated clearly and coherently the specific requirements for assessing and certifying the authenticity of electronic records, and produced a richer understanding of the relationships between the archival and the digital properties of the records. In its process models of selecting and preserving electronic records, the Interpares project has delineated the decision processes involved in these activities, providing archives and other institutions responsible for digital preservation a firmer and broader basis for evaluating and selecting information technology products on the basis of archival criteria.[3]

The problem of rapidly growing quantities of electronic records demanded a frontal assault in its own right. It combined an engineering problem of scaling input/output operations to cope with enormous numbers of physical files with the archival challenge of identifying and controlling the records and aggregates of records contained in those files. NARA attacked this problem by joining the distributed object computation testbed, a partnership of the Defence Advanced Research Projects Agency, the US Patent and Trademark Office, with the San Diego Supercomputer Centre (SDSC) at the University of California, San Diego, as the leading research centre. The specific challenge NARA added to this testbed was to find a way to preserve exponentially growing quantities of electronic records while respecting the archival principles of provenance and original order, and without raising unnecessary barriers to access to the records.[4]

This project took those initial constraints, of respecting provenance and original order and optimising accessibility, and recast them as the guiding principles of a comprehensive method for digital preservation, initially termed ‘persistent object preservation.’ This method reverses the focus that had characterised previous efforts to develop techniques for digital preservation. Techniques like emulation and migration focus on overcoming technological obsolescence. In contrast, persistent object preservation focuses on the essential properties of the objects that are to be preserved and insulates those properties from the effects of continuing change in information technology. For archives, the objects to be preserved are records and archival aggregates of records, rather than data types. Options for preservation should be selected on the basis that they maintain the content and documentary form of individual records and the organic relationships among records in files, series, and archival fonds. Given that most electronic records and aggregates are created in data types which are subject to obsolescence, the persistent object method entails transforming the digital objects into formats that are more suited for long-term maintenance and access. Superficially, this looks like migration, but there is a basic difference. Migration techniques typically move objects from older to newer formats, where the control on the process is at the level of data types. But data types do not necessarily correlate with the essential characteristics of records. Hence, migration can only guarantee that old data remains accessible in new technology. It cannot guarantee that records remain authentic. Rather than seeking to keep old software working or translating stored information to new data formats, the persistent object method makes the essential features of the objects to be preserved explicit in formal models and encapsulates the objects in metadata defined in those models, thus differentiating between data (encapsulated objects stored as bit sequences) and information (syntactic and semantic relationships among data articulated in models). It enables systems to evolve independently at the data and information levels by using software mediators between them. Similarly, a system can be designed to provide the basic functions of ingest, management, and dissemination, prescribed by the OAIS model, using software mediators to make each function relatively independent of the others. Thus, over time, the hardware and software used to implement any function, at any level, can be replaced without requiring changes in other parts of the architecture or in the collections being preserved. Only the mediators that provide the interfaces need to be adjusted in response to such change.[5]

The relationship between NARA staff and the scientists and engineers at SDSC developed into a mutually reinforcing collaboration. The results of successive rounds of research defined for archivists unprecedented possibilities for approaching the challenge of preserving electronic records from a systematic, open-ended perspective. Conversely, as the computer scientists and engineers conducting the research progressively enriched their understanding of the archival challenge, they were articulating a comprehensive, knowledge-based information management architecture for digital preservation which not only incorporated the full range of archival requirements from accession to dissemination and from individual records to archival fonds, but also offered clear potential for application in other fields, including not only digital libraries, but also research in natural sciences such as astronomy, physics, and neurosciences.[6]

[2] Consultative Committee for Space Data Systems, ‘Reference model for an open archival information system, CCSDS 650.0-R-2, *Red book*, July 2001. The model has been adopted and is being published by the ISO. Available at: http://www.ccsds.org/documents/pdf/CCSDS-650.0-R-2.pdf

[3] http://www.interpares.org/index.htm

[4] http://www.darpa.mil/ipto/psum1999/d642-0.html
See also: http://www.sdsc.edu/DOCT/

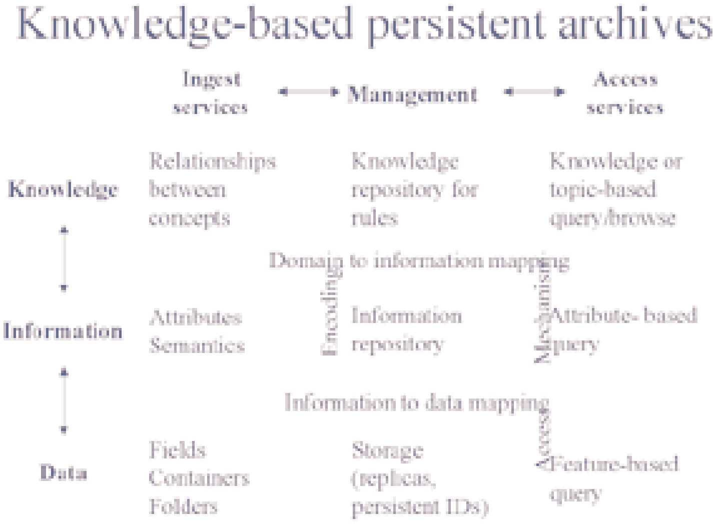
[5] Thibodeau, K., Moore, R., Baru, C., ‘Persistent object preservation: advanced computing infrastructure for digital preservation’, *DLM-Forum — European citizens and electronic information: the memory of the information society, cooperation Europe-wide*, Brussels, 18–19 October 1999, pp. 113–119, available at: http://europa.eu.int/ISPO/dlm/fulltext/full_thib_en.htm

[6] Rajasekar, A., Moore, R., ‘Data and metadata collections for scientific applications’, European high performance computing conference, Amsterdam, Holland, 26 June 2001, available at: http://www.npaci.edu/DICE/Pubs/Data-management_moore.pdf

The fertility of this interaction between archivists, on the one hand, and computer scientists and engineers, on the other, is most evident in the maturation of the initial 'persistent object' preservation method into the 'persistent archives' architecture. The persistent archives concept adds a knowledge layer to the architecture, on top of the data and information layers. This addition is at once simple in its basic formulation and prolific in its implications. At bottom, all digital information is stored as data (byte sequences). These byte sequences can only be processed and interpreted correctly given syntactic and semantic context at the information level. Information, then, is defined as valid and meaningful combinations of data. In turn, information can only be understood and exploited in the context of domain-specific knowledge. Knowledge, then, is defined as combinations of pieces of information or knowledge.[7] The following figure depicts the persistent archives architecture. It is constructed of three rows corresponding to the knowledge, information and data layers, from the top down, and three columns corresponding to the OAIS functions of ingest, maintenance and dissemination, from left to right. The principal entities or processes involved in each function at each level are shown in the individual cells, and the types of mediation services are indicated between the rows and columns.

To illustrate the differentiation over levels, consider the types of queries that could be executed. At the data level, regardless of how information or knowledge are managed, simple queries such as string searches for character data or pattern matching for other data types could be executed against the stored data. The process would be similar to searches run on the world wide web today. If data were stored with tags that identified attributes, one could incorporate this additional information in slightly more sophisticated searches, such as where tag = 'income',or where tag = 'author' and the data value = 'John Mark'. Moving up to the information level would enable us to apply syntactic and semantic criteria. For example, where a data model indicated that a database contained elements for year, location, earned income, and occupation, one could execute queries asking for the income of electronics engineers in a specified city in 1999. If the system could materialise the structure of the database as indicated in the data model, and populate that structure with the stored data, it would not matter whether these data elements were stored in the same flat file or were scattered in different tables of a complex relational database. In order to reinstantiate the database on a target platform at some time in the future, the system would need a software mediator which could translate the metadata stored in the information repository into terms recognisable by the target technology. The maintenance of the logical model apart from the stored data would provide independence between information management and data storage. The software mediator would achieve independence between the management of information about stored objects and collections and access to the preserved records.

Providing an appropriate response to a question such as, 'Find all diplomatic correspondence concerning the North-American Free Trade Agreement', would require greater sophistication than the examples cited so far. For example, a string search of the web on the phrase, 'North-American Free Trade Agreement', identified over 160 000 relevant items. Adding the phrase 'diplomatic corre-


Knowledge-based persistent archives
Ingest services
Management
Access services
Knowledge
Relationships between concepts
Knowledge repository for rules
Knowledge or topic-based query/browse
Domain to information mapping
Encoding
Mechanisms
Information
Attributes Semantics
Information repository
Attribute- based query
Information to data mapping
Access
Data
Fields Containers Folders
Storage (replicas, persistent IDs)
Feature-based query

[7] Ludäscher, B., Marciano, R., Moore, R., 'Preservation of digital data with self-validating, self-instantiating knowledge-based archive', *ACM Sigmod Record*, 2001, 30(3), pp. 54–63.

spondence' reduced the result set to six items, but none of them contained any diplomatic correspondence. An appropriate response to this type of query would benefit from the application of knowledge from several different domains, including foreign affairs: that NAFTA is a treaty among Canada, Mexico, and the United States; government: which agencies of those three governments are likely to engage in diplomacy regarding trade; archival science: what types of records comprise correspondence; what distinguishes diplomatic from other correspondence, and where in the record-keeping systems of the relevant governmental entities would diplomatic correspondence be found. To execute such a query, an archival system would need a knowledge base for topics, such as 'North America' and 'free trade' and another for rules, such as those concerning the conduct of foreign affairs in the three nations and those concerning record-keeping in the government entities. It would also entail mapping the relevant topics and rules to objects and the information level and from there to the data level. Ensuring that the outcome of the query was access to authentic copies of the relevant correspondence would also require application of knowledge about archival requirements for preservation to the information about the management of information about the records, the data that constitute the records, and the processes for maintaining and reproducing them. Separating knowledge management from information management has obvious benefits. It enables tracking of significant changes in records systems over time. For example, diplomatic correspondence within the US Department of State was moved from paper to digital form in the 1970s, and in the intervening decades the electronic system underwent several migrations. Some of the technology changes entailed significant changes in the records, although such changes might only be implicit or not apparent to someone reading only one message at a time. A knowledge base can enable such changes to be made explicit and readily avialable. But it would also enable maintaining coherency across lower-level variability. For example, each of the State Department systems would be represented at the information level by a different schema; however, the diplomatic messages constitute a single records series, one that has been in continuous existence for over a century. A knowledge level concept of the series enables it to be managed as a single archival aggregate, but also makes it possible to translate a query to execute appropriately against difference instantiations of the records system.

The persistent archives approach reflects the conclusion of technologists that it is impossible to satisfy archival requirements for preservation and access to electronic records without systematic infusion of archival knowledge. This domain knowledge is essential to distinguish archives of records from other types of information collections in order to ensure records and archival aggregates of records. There are several areas where archival knowledge needs to be applied. Most obvious is the need to ensure the continuing value of records as evidence of the activities in which they are produced and used. This entails application of the archival principles of provenance and respect for original order. In addition, archival knowledge of requirements for preservation dictates what information is required to assess and assert the authenticity of the preserved records.

Domain knowledge is also needed to understand records. Unlike other types of information objects, such as publications, records are created within a universe of discourse where there is often a high degree of shared information and expectations among participants. This common knowledge includes both specific empirical information about prior steps in a multi-step process, generic knowledge about the process, and expectations about both subsequent steps and the norms for recording and communicating information about the process. In such contexts, important information is often conveyed by form, as well as by substance. In processes involving businesses and government agencies, for example, participants expect certain forms to be used for certain types of transactions, and certain signatures to be required for authorisation of certain types of actions. Common knowledge enables processes to be carried out efficiently and provides a systemic check on their validity and on the reliability of their records. For example, participants in a recurring process are likely to note deviations from prescribed documentary forms or absences of required authorisations. To enable parties who were not participants in a process to understand the records of that activity, often long after the fact, an archival system should contain and convey information about the types of records typically produced, the elements of intrinsic and extrinsic form of each type, the relationships between processes and records, and also the implied knowledge that was either common to participants or would have resulted from recognised discrepancies between generic knowledge and expectations on the one hand and specific instances on the other.

The persistent archives architecture came to fruition thanks to a new partnership NARA formed in March 2000 with the US National Science Foundation's National Partnership for Advanced Computational Infrastructure.[8] This partnership offered the tantalising prospect of simultaneously fine

[8] National Archives and Records Administration, 'National archives announces plan for collaboration with National Science Foundation to create an electronic records archives', press release, 28 March 2000, available at: http://www.archives.gov/media_desk/press_releases/nr00-58.html

tuning the preservation method to address specific archival requirements while finding solutions capable of addressing these requirements in mainstream technologies being developed to enable electronic commerce and electronic government. Archives and advanced computation are, in one sense, at opposite ends of an intellectual spectrum: the one seeks to preserve the past the other to create the future. But it would be more accurate to say that, while they focus on different axes, they both face a common problem. To preserve electronic records, archives need in effect to develop interoperability among disparate systems, chronologically spanning generations of information technology. To enable e-commerce, e-government, scientific research, and education in the digital era, those responsible for developing information and communications technologies must provide for interoperability across disparate administrative and political domains, simultaneously spanning several generations of technologies implemented in these domains at any given time. The common need for interoperability in turn entails a need for persistence of the information assets created and used in these domains. These assets have to be carried — intact and authentic — across space, time, technologies, and human (political, institutional, social and cultural) domains. These common interests have generated a sustained and growing partnership between NARA and NSF, and a widening range of research initiatives at the University of California, Berkeley, the University of Maryland, Ohio State University, and the University of Urbino, Italy, as well as at SDSC. Further expansion in the range of active researchers is expected in the near future.[9]

The common interests of archives and institutions promoting advanced technologies are especially important in the government arena. By its nature, a democratic government is responsible to its citizens. Records are an essential instrument of that accountability. If government archives — at national and other levels — cannot find ways to preserve and deliver authentic electronic records over time, governments will not be able to fulfill a fundamental responsibility. The importance of this challenge has been recognised at the highest levels of the US Government. Each year, the President sends to the Congress a supplement to the President's budget, which sets out the overall picture of the government's support for research and development in computer and networking technologies. In the Supplement for the fiscal year 2002, the Interagency Working Group on Information Technology Research and Development identified 'managing and enabling worlds of knowledge' as one of the major challenges facing the nation. Within this context, it asserted, 'Strategies to assure long-term preservation of digital records constitute another particularly pressing issue for research. As storage technologies evolve with increasing speed to cope with the growing demand for storage space, the obsolescence of older storage hardware and software threatens to cut us off from the electronically stored past.' It also specified, 'How to determine, collect, and preserve what is of value in the world's dizzying new digital output now joins older questions of how and what to digitise from humanity's pre-digital knowledge stores as issues for archivists.'[10] These assertions are not simply abstract statements of principle. The Congress translated this direction into real support for the preservation of electronic records in the NARA appropriation for fiscal year 2002. The current NARA budget includes USD 22 302 000 for preserving electronic records. While some of this money is being used for current operations, the majority is devoted to undertaking the development of a new archival system for electronic records. According to John Carlin, the Archivist of the United States, the Electronic Records Archives 'is NARA's strategic response to the challenge of preserving, managing, and accessing electronic records. Our goal is to build a nationwide digital archive that preserves and provides access to virtually any type of electronic record created anywhere in the Federal Government at any time.'[11]

[9] Thibodeau, K., 'Building the archives of the future — advances in preserving electronic records at the National Archives and Records Administration', D-Lib Magazine, February 2001, Volume 7, Number 2. http://www.dlib.org/dlib/february01/thibodeau/02thibodeau.html

[10] National Science and Technology Council, 'Networking and information technology research and development', Supplement to the President's Budget For FY 2002, a report by the Interagency Working Group on Information Technology Research and Development, Washington, July 2001, pp. 22–23, available at: http://www.ccic.gov/pubs/blue02/index.html

[11] National Archives and Records Administration, 'ERA vision statement', NARA notice: NARA 2002-135, 10 May 2002.

Progresos realizados en la administración americana por lo que respecta al conocimiento y a la acción en el ámbito de la conservación digital

Keneth Thibodeau

Desde 1998, la Administración Nacional de Archivos y Documentos (NARA) de Estados Unidos ha patrocinado y colaborado en varias iniciativas relativas a diversas cuestiones, desde infraestructura informática avanzada hasta los requisitos para garantizar la autenticidad de los archivos. Estos proyectos han supuesto importantes asociaciones con otros organismos del Gobierno estadounidense, tales como National Science Foundation, Defense Advanced Research Projects Agency y Army Research Laboratory; con universidades y organismos de investigación, incluidas las universidades de California, Maryland y Urbino y el Georgia Tech Research Institute, y colaboraciones multinacionales y multidisciplinares tales como el proyecto InterPARES y la Global Electronic Records Association. Esta diversidad de iniciativas se ha centrado en los objetivos básicos de desarrollar los conocimientos necesarios para comprender los requisitos y evaluar las opciones para la conservación a largo plazo de los documentos electrónicos y el acceso continuo a ellos, así como de fomentar el desarrollo o la transferencia de las tecnologías necesarias para realizar con éxito la conservación y el acceso.

Los resultados de estas actividades son los siguientes: 1) un cambio fundamental en el concepto de los retos de la conservación y del acceso continuo a la información digital; 2) grandes progresos en las áreas de la informática y de la ingeniería, que son críticas para tener éxito al abordar estos retos; 3) la concentración de la dirección estratégica de la NARA en construir soluciones sólidas para la conservación y el acceso a los documentos electrónicos, y 4) grandes beneficios en cuanto al reconocimiento de la importancia y el valor de estos esfuerzos y su aplicabilidad a una amplia gama de importantes actividades.

Nuestro concepto de los retos relativos a la conservación de los archivos electrónicos ha variado desde el limitado objetivo de encontrar medios para superar problemas tecnológicos tales como la obsolescencia del *hardware* y del *software* y la fragilidad de los medios digitales, hasta la articulación de un modelo abstracto en el que la conservación y el acceso continuo se consideran íntimamente ligados entre sí y con la infraestructura de la tecnología de la información necesaria para apoyar el comercio electrónico y la administración electrónica.

Esta visión conceptual se ha articulado paralelamente al progreso del desarrollo de una arquitectura de gestión de la información para archivos permanentes basada en el principio de que la conservación de la información digital debe ser esencialmente independiente de la tecnología de la información específica utilizada para aplicar la solución. Este principio de independencia de la infraestructura tiene dos facetas esenciales. En primer lugar, aísla los objetos digitales y cualquier recogida arbitraria de objetos digitales, incluso en el nivel de los fondos de archivo, de las vicisitudes de los cambios tecnológicos, a fin de reforzar la autenticidad continua de la información conservada. En segundo lugar, permite que las soluciones se adapten de forma dinámica, tanto para abarcar nuevos tipos de objetos digitales que puedan surgir en el futuro como para poder hacer frente a las demandas de envío de los objetos conservados, realizadas por los usuarios, a través de una variedad grande y cambiante de plataformas informáticas. La arquitectura de los archivos persistentes se ha desarrollado hasta el punto de incluir como componente básico esencial la introducción del conocimiento temático sobre los datos conservados. La necesidad y el valor de incluir el conocimiento del dominio se han demostrado en casos que van desde los registros de la actividad legislativa en el Congreso estadounidense hasta los datos experimentales de las neurociencias.

Hace unos años, frente a los retos desalentadores y crecientes que planteaban los registros electrónicos, la NARA sólo podía articular una estrategia consistente en dividir los problemas y resolverlos uno por uno. Los proyectos elaborados a finales de los años noventa clasificaron las dificultades en función de las categorías de aplicaciones y los tipos de datos, con un plan subsidiario para abordar cada categoría, esencialmente en forma secuencial. En la actualidad, la arquitectura de los archivos permanentes proporciona la base para una estrategia completa y coherente en la que diferentes categorías de aplicaciones se consideran casos especiales que suponen diferencias marginales, y no básicas, en cuanto a las soluciones. De conformidad con esta estrategia, la NARA ha puesto en marcha el programa de documentos y archivos electrónicos (Electronic Records Archives Program) para construir los archivos del futuro, previstos como un sistema federado distribuido por Internet y capaz de absorber, conservar y expedir virtualmente cualquier tipo de archivos electrónicos.

El éxito de estos esfuerzos se ha reconocido ampliamente, de forma significativa para la NARA, en los más altos niveles de la administración americana. El Gabinete del Presidente, en su plan gubernamental de investigación y desarrollo en el ámbito de las tecnologías de información y la creación de redes, considera la conservación digital en general y la selección y conservación de archivos en particular como elementos importantes para "facilitar la emergencia de mundos de conocimiento". En el presupuesto actual, el Congreso estadounidense ha aprobado la petición de la Casa Blanca de mayor financiación, asignando más de 2 millones de dólares a apoyar los esfuerzos de la NARA para hacer frente a los retos en el ámbito de los archivos electrónicos.

Kenntnisse und Maßnahmen im Bereich digitale Archivierung: der aktuelle Stand bei der US-Regierung

Keneth Thibodeau

Seit 1998 finanziert und beteiligt sich die National Archives and Records Administration (NARA) der Vereinigten Staaten an einer Reihe von Initiativen zur Erforschung der verschiedensten Fragen von der modernen EDV-Infrastruktur bis hin zu Authentizitätsanforderungen bei der Archivierung. Diese Projekte erfolgen auf der Basis einer breit angelegten Partnerschaft mit anderen Stellen der US-Regierung wie der National Science Foundation, der zentralen FuE-Einrichtung des Verteidigungsministeriums (DARPA) und dem Forschungslabor der US-Army, mit Universitäten und Forschungseinrichtungen wie den Universities of California, Maryland und Urbino und dem Georgia Tech Research Institute, sowie im Rahmen multinationaler und multidisziplinärer Vorhaben wie dem Projekt InterPARES und der Global Electronic Records Association. Im Mittelpunkt dieser breiten Palette von Initiativen stehen folgende Kernziele: Erweiterung der erforderlichen Kenntnisse über die Anforderungen der Langzeitarchivierung elektronischer Unterlagen und ihrer ständigen Zugänglichkeit sowie die Bewertung entsprechender Optionen, und Förderung der Entwicklung oder des Transfers der für die Archivierung und den Zugang notwendigen Technologien.

Die bisherigen Ergebnisse dieser Aktivitäten lassen sich folgendermaßen zusammenfassen: (1) ein grundsätzliches Umdenken im Hinblick auf die Aufgabenstellungen der Konservierung und der ständigen Zugänglichkeit von digitalen Informationen, (2) wesentliche Fortschritte im Bereich Informatik, die für Erfolge bei der Erfüllung dieser Aufgabenstellungen entscheidend sind, (3) strategische Ausrichtung der NARA auf die Erarbeitung tragfähiger Lösungen für die Konservierung und Zugänglichkeit elektronischer Unterlagen und (4) große Fortschritte bei der Anerkennung der Bedeutung und des Werts dieser Bemühungen sowie deren Anwendbarkeit auf ein breites Spektrum von Aktivitäten.

Unsere Vorstellung von den mit der Archivierung elektronischer Unterlagen verbundenen Herausforderungen hat sich verlagert und richtet sich nicht mehr allein auf die Suche nach Mitteln zur Überwindung technischer Probleme wie dem raschen Veralten von Hard- und Software und der Unsicherheit digitaler Datenträger. Vielmehr geht es jetzt um die Formulierung eines abstrakten Modells, bei dem Konservierung und ständiger Zugang als untrennbar miteinander verknüpft und eingebunden in die informationstechnische Infrastruktur betrachtet werden, die zur Unterstützung von e-Commerce und e-Government erforderlich ist.

Diese neue Vorstellung entwickelte sich bei gleichzeitigen Fortschritten bei der Herausbildung einer Informationsmanagement-Architektur für dauerhafte Archive, die auf dem Grundsatz beruht, dass die Konservierung digitaler Informationen im Wesentlichen unabhängig von der zur Implementierung verwandten konkreten Informationstechnologie erfolgen muss. Dabei spielen zwei maßgebliche Gesichtspunkte eine Rolle. Zum einen werden die digitalen Objekte und jede willkürliche Sammlung digitaler Objekte – bis hin zur Ebene von Archivbeständen – von den Wechselfällen des technologischen Wandels isoliert, um für die anhaltende Authentizität der konservierten Informationen zu sorgen. Zum anderen lassen sich auf diese Weise Lösungen dynamisch so anpassen, dass neue Arten digitaler Objekte und Sammlungen, die in der Zukunft entstehen können, aufgenommen werden können und dass den Forderungen der Benutzer nach Ausgabe der konservierten Objekte auf den verschiedensten und sich verändernden Rechnerplattformen Rechnung getragen wird. Die Architektur dauerhafter Archive hat sich so weit entwickelt, dass nunmehr Domainkenntnisse zu den konservierten Sammlungen als wesentlicher Bestandteil mit enthalten sind. Notwendigkeit und Wert der Aufnahme von Domainkenntnissen wurden an konkreten Fällen nachgewiesen, die von Unterlagen über gesetzgeberische Tätigkeiten des Kongresses bis hin zu Experimentaldaten der Neurowissenschaften reichen.

Vor ein paar Jahren konnte die NARA angesichts der gewaltigen und rasant zunehmenden Herausforderungen, die elektronische Unterlagen mit sich brachten, lediglich eine Strategie des „Teile und herrsche“ formulieren. Ende der 90er Jahre erstellte Pläne teilten die Herausforderungen nach Anwendungskategorie und Datentyp, wobei anhand von Teilplänen die einzelnen Kategorien im Wesentlichen der Reihenfolge nach zu behandeln waren. Heute bildet die Architektur für dauerhafte Archive die Grundlage für eine umfassende und einheitliche Strategie, bei der unterschiedliche Anwendungskategorien als Spezialfälle betrachtet werden, die geringfügig, aber nicht grundsätzlich unterschiedlicher Lösungen bedürfen. Entsprechend dieser Strategie hat die NARA ein Elektronisches Archivprogramm aufgelegt, das dem Aufbau eines Archivs der Zukunft in Form eines über das Internet verteilten föderalen Systems dient, mit dem praktisch alle Arten von elektronischen Aufzeichnungen aufgenommen, konserviert und ausgegeben werden können.

Der Erfolg dieser Bemühungen findet breite Anerkennung, vor allem auch auf den höchsten Ebenen der amerikanischen Regierung, was für die NARA von großer Bedeutung ist. In dem alle Regierungsstellen umfassenden Plan für Forschung und Entwicklung im Bereich Informations- und Netzwerktechnik hat das Präsidialamt die digitale Archivierung im Allgemeinen und die archivische Bewertung und Konservierung von Unterlagen im Besonderen als wichtige Elemente beim „Aufbau von Welten des Wissens“ anerkannt. Im laufenden Haushalt hat der Kongress dem Antrag des Weißen Hauses für eine erhöhte Mittelausstattung stattgegeben und mehr als 2 Mio. US-Dollar für die Erfüllung der Aufgaben im Bereich elektronische Unterlagen durch die NARA bereitgestellt.

Sur la connaissance et l'action dans le domaine de la conservation numérique: le point sur les progrès réalisés dans l'administration américaine

Keneth Thibodeau

La NARA (National Archives and Records Administration) des États-Unis parraine et collabore à un certain nombre d'initiatives depuis 1998 qui se proposent d'analyser diverses questions, de l'infrastructure informatique avancée aux spécifications requises pour l'authenticité des archives. Ces projets ont impliqué de larges partenariats avec d'autres organismes de l'administration américaine, notamment avec la fondation NSF (National Science Foundation), l'agence Defense Advanced Research Projects Agency, l'Army Research Laboratory, avec des universités et des instituts de recherche, parmi lesquels les universités de Californie, du Mariland et d'Urbino et le Georgia Tech Research Institute, ainsi que des collaborations multinationales et pluridisciplinaires, à l'instar des projets InterPARES et Global Electronic Records Association. Ces initiatives ont ciblé des objectifs de base: faire progresser l'état des connaissances nécessaires pour connaître les conditions requises de la conservation à long terme des documents électroniques et évaluer les solutions possibles, assurer un accès durable à ces archives et encourager le développement ou le transfert des technologies nécessaires pour réussir la conservation des archives et leur accessibilité.

Toutes ces activités ont débouché sur les résultats cumulés suivants: 1) un changement fondamental dans la conception des défis que posent la conservation et un accès durable à l'information numérique; 2) des progrès substantiels dans le domaine de la science et du génie informatiques, qui sont décisifs pour réussir à relever ces défis; 3) une orientation stratégique de la NARA, qui cible la mise en place de solutions robustes pour la conservation des documents électroniques et leur accessibilité et 4) des avancées majeures dans la reconnaissance de l'importance et de la valeur des efforts déployés et dans leurs possibilités d'application à un large éventail d'autres activités.

Notre conception des défis concernant la conservation des archives électroniques a évolué d'une vision étroite, qui consistait à trouver des moyens pour surmonter des problèmes technologiques comme l'obsolescence des matériels et des logiciels et la fragilité des supports numériques, vers la formulation d'un modèle abstrait, dans lequel la conservation et l'accès permanent sont vus comme inextricablement liés l'un à l'autre et comme intimement mêlés aux technologies de l'information nécessaires pour supporter le commerce et l'administration électroniques.

Cette vue conceptuelle s'est trouvée formulée parallèlement aux progrès réalisés dans le développement d'une architecture de gestion documentaire destinée aux archives permanentes, reposant sur le principe selon lequel la conservation d'informations numériques doit être pour l'essentiel indépendante des technologies de l'information spécifiques utilisées pour implémenter la solution. Ce principe de l'indépendance de l'infrastructure revêt deux aspects primordiaux: tout d'abord, il protège les objets numériques et toute autre collection arbitraire d'objets numériques — voire un fonds d'archives — des caprices du changement technologique afin d'assurer l'authenticité constante de l'information conservée. Ensuite, il permet une adaptation dynamique des solutions à la fois pour qu'elles puissent s'étendre aux nouveaux types d'objets numériques et de collections susceptibles d'émerger à terme et pour qu'elles répondent aux demandes des chercheurs souhaitant que les objets conservés leur soient communiqués sur un large éventail de plates-formes informatiques. L'architecture des archives permanentes a évolué jusqu'au stade où elle inclut des connaissances thématiques sur les collections archivées en tant que composante essentielle. La nécessité et l'importance de cette incorporation des connaissances thématiques ont été démontrées dans de nombreux exemples, qui vont des archives de l'activité législative du Congrès américain aux données expérimentales des neurosciences.

Face aux difficultés décourageantes et croissantes que posaient les archives électroniques il y a quelques années, la NARA ne pouvait que formuler une stratégie consistant à scinder les problèmes et à les résoudre un à un. Les projets élaborés à la fin des années 90 classaient les difficultés en fonction des catégories d'applications et des types de données, avec un projet subsidiaire permettant d'aborder chaque catégorie, essentiellement sous forme séquentielle. Aujourd'hui, l'architecture des archives permanentes jette la base d'une stratégie globale et cohérente, dans laquelle différentes catégories d'applications sont perçues comme des cas particuliers, entraînant des différences marginales, et non fondamentales, dans les solutions. Conformément à cette stratégie, la NARA a lancé le programme en faveur des documents et des archives électroniques afin de mettre en place les archives du futur, vues comme un système fédéré réparti sur l'internet et capable d'admettre, de conserver et de mettre à disposition presque n'importe quel type d'archives électroniques.

Le succès de ces efforts est largement reconnu, même aux niveaux les plus élevés de l'administration américaine, ce qui est encore plus important pour la NARA. Dans son plan gouvernemental pour la recherche et le développement dans les technologies de l'information et des réseaux, le cabinet du président a reconnu que la conservation numérique en règle générale, et le tri et la conservation des archives en particulier, constituaient des éléments majeurs pour "faciliter l'émergence des mondes de la connaissance". Dans le budget actuel, le Congrès américain a accédé à la demande d'augmentation de budget de la Maison blanche, et a ainsi fourni plus de 2 millions de dollars pour soutenir les initiatives de la NARA dans le domaine des archives électroniques.