Desarrollo de una política de preservación digital: tecnología, planificación y perseverancia.

Alejandro Bia y Manuel Sánchez

Biblioteca Virtual Miguel de Cervantes,
Universidad de Alicante,
Apdo. de correos 99, E-03080, Alicante, España
{alex.bia, manuel.sanchez}@cervantesvirtual.com
http://cervantesvirtual.com/

Resumen. La preservación digital es una batalla perdida a largo plazo, ¿o no es así? La ley de la entropía juega en contra de este propósito. No importa el medio físico que elijamos: éste se degradará con el paso de tiempo con la consiguiente pérdida de información. El objetivo de la preservación digital es retardar esta degradación tanto como sea posible. Las técnicas usuales de preservación digital requieren de un esfuerzo periódico y planificado. La preservación digital, mediante una política adecuada de renovación de la información grabada y actualización de los formatos de datos, puede garantizar una vida muy larga a los recursos digitales.

Este trabajo trata tanto de los aspectos técnicos relativos a la preservación digital, como son la obsolescencia y envejecimiento de los soportes físicos, el rejuvenecimiento de la información digital, la conversión de formatos y los sistemas de búsqueda y recuperación de recursos almacenados, así como de los aspectos metodológicos y de definición de una política de preservación adecuada que establezca las reglas a seguir y defina qué es lo que debe ser preservado y lo que no.

Según P. L. Shillingsburg [1] "el editor que use un sistema de codificación universal para desarrollar una edición electrónica con una aplicación multiplataforma habrá creado una herramienta disponible para cualquier lector con acceso a un ordenador y habrá asegurado la longevidad de ese trabajo editorial para las generaciones de software y hardware venideras". Estas ideas pueden extenderse a todo tipo de recurso multimedia utilizable por las bibliotecas digitales. En la Biblioteca Virtual Miguel de Cervantes, estamos convencidos de la importancia y el interés de preservar estos esfuerzos editoriales para las generaciones de lectores venideras.

"La cultura, cualquier cultura ... depende de la calidad de sus registros de conocimiento" [2].

"Una biblioteca digital es la combinación de una colección de objetos digitales (repositorio); las descripciones de esos objetos (metadatos); un conjunto de usuarios (clientes o público o usuarios); y sistemas que ofrecen una serie de servicios como la captura, indexación, catalogación, búsqueda, navegación, recuperación, entrega, archivo y preservación." [3]

1 ¿Qué es la preservación digital?

Según Levy [4], la preservación de la información digital es un problema difícil y no muy bien entendido. Mientras que la información en papel y en otros soportes duraderos (por ejemplo el microfilm) puede durar cientos, y en algunos casos miles de años, la información codificada en formato digital es poco probable que dure más de una década o dos. La razones de esto son bien conocidas: el envejecimiento de los soportes digitales, la obsolescencia de los formatos, del software y del hardware, así como la falta de compatibilidad hacia atrás de los nuevos sistemas.

Oya Rieger ve la preservación digital como un problema organizativo [5], y sin duda, las políticas empresariales respecto a la preservación digital son fundamentales para el éxito de cualquier iniciativa al respecto. Según ella, el problema son los elementos e infraestructuras tecnológicos y empresariales rápidamente cambiantes, y el objetivo es mantener la habilidad de localizar, desplegar y usar las colecciones digitales. En su opinión, la preservación digital son las actividades de gerencia que aseguran el acceso continuo a los materiales digitales frente a los elementos tecnológicos y empresariales rápidamente cambiantes [6].

¿La preservación digital es la preservación de los originales por métodos digitales, o la preservación de los propios materiales digitales? Se puede entender por *preservación digital* la preservación de los artefactos físicos mediante su digitalización, pero también la preservación de los propios recursos digitales. La digitalización de materiales originales valiosos es, a menudo, realizada con dos propósitos: brindar un mejor acceso a esos recursos mediante copias digitales, y mejorar la preservación de los originales [7]. Sin embargo, la preservación es a menudo un fin secundario, preservándose solamente los originales mediante la reducción del acceso físico a los mismos. La versión digital generalmente no es considerada como un recurso de preservación primario. Por importante que sea la preservación de los originales a través de la digitalización, probablemente estos no estén tan en riesgo inmediato como los propios datos y recursos digitales, siendo los recursos que se crean digitalmente y no como copia de recursos físicos los que están más en peligro [8]. En este artículo nos referiremos exclusivamente a la preservación digital, es decir, a la que trata de mantener utilizables e inalterados los recursos digitales a través del tiempo.

2 Diferencias entre preservación y copias de seguridad

La preservación digital es diferente de las copias seguridad. En primer lugar no es lo mismo lo que se ha de preservar que lo que se ha de guardar como copia de seguridad. Las copias de seguridad son una protección contra eventos catastróficos (rotura de un disco o pérdida de datos por apagones, por ejemplo). Lo que se guarda como copia de seguridad en una biblioteca digital son, básicamente, dos cosas: por un lado la información publicada en el servidor (recursos digitales más información de catálogo) y, por otro lado, los recursos digitales en proceso de edición. La preservación digital sin embargo, no se ocupa de respaldar ni los

datos del servidor ni el material de trabajo diario, sino de salvaguardar los recursos digitales que necesitaremos en el futuro. Un ejemplo claro de esta diferencia lo tenemos en el caso de las imágenes gráficas: nos interesa hacer copia de seguridad de las imágenes JPG (comprimidas con pérdida de calidad) que colgamos en el servidor, pero sin embargo nos interesa hacer copias de preservación de las imágenes TIFF de alta calidad que han dado lugar a las JPG pero que no publicamos por su mayor tamaño y lentitud de transmisión. En otras palabras, en el servidor nos interesa usar imágenes comprimidas de menor calidad pero que tarden menos en transmitirse, mientras que nos interesa preservar imágenes de la mayor calidad posible para usos futuros como puede ser el caso de la producción de CD-ROMs que no poseen las mismas limitaciones de velocidad que la Internet, o para el uso en redes de mayor ancho de banda que la Internet actual. Alexa McCray y otros [9] hacen una separación entre el material para archivo y los derivados para acceso público. Su modelo de biblioteca digital incluye una versión maestra de la biblioteca digital con los recursos de alta calidad (los que se preservan) y una biblioteca de acceso público con formatos generados automáticamente a partir de la primera.

Si bien las copias de seguridad, al igual que las de preservación, se basan en la redundancia de la información mediante grabaciones periódicas, ni la forma de organizar esta grabaciones ni los tiempos son los mismos. Las copias de seguridad pueden seguir diversos métodos conocidos: copia integral, copia incremental o copias rotativas, por ejemplo, y la periodicidad generalmente es alta (diaria o semanal). En el caso de la copias de preservación, por el contrario, el método suele ser la grabación integral del material una vez y el copiado del mismo una vez al año o cada año y medio en otro soporte nuevo (rejuvenecimiento).

En ambos casos, se utilizan mecanismos de control de la integridad de los datos al momento de hacer las copias, mediante algoritmos de redundancia que verifican que los datos se mantienen tal como han sido grabados.

3 Definición de políticas de preservación

Se debe establecer la preservación digital como una responsabilidad institucional con un firme soporte financiero, apoyo a nivel gerencial y un compromiso de todo el personal. La definición de un plan de preservación pasa por dar respuesta a las siguientes preguntas:

1. ¿Qué guardar y porqué guardarlo? [4]
2. ¿Dónde guardarlo?
3. ¿Hasta cuándo guardarlo?
4. ¿Cómo encontrarlo después?
5. ¿Cómo hacer que se mantenga inalterado?
6. ¿Cómo evitar que se vuelva obsoleto?

En primer lugar, se deben seleccionar y crear colecciones digitales con un valor duradero. Luego debe haber una política de preservación bien definida, que establezca las reglas y procedimientos a seguir, así como lo que debe ser

preservado. Esta política debe ser revisada periódicamente tanto para mejorar los métodos como para redefinir el conjunto de objetos a ser preservados. A los objetos preservados se les debe asignar un límite de vida. Algunos serán más perecederos que otros, y estas duraciones deberían ser revisadas periódicamente.

3.1 Integridad digital

Algunas de las posibles causas de defectos en la información digital o pérdida de datos son: errores de gestión y negligencia, fallos técnicos y mecánicos, errores del operador, virus, cambios no autorizados y no documentados, obsolescencia o incompatibilidad del software, pérdida de programas, metadatos incompletos, envejecimiento de la información.

3.2 Medios de almacenamiento digital

Por otro lado, ninguno de los medios digitales de hoy en día garantiza la longevidad de los información. Los medios magnéticos tienen una vida sorprendentemente corta. Los discos compactos son más estables pero no se puede predecir su duración, que depende en buena medida de la calidad de los mismos. Si no se toman medidas gran parte de la información se perderá en unas pocas de décadas.

3.3 Procedimientos de preservación

Según Oya Rieger [6], se debe:

- Almacenar los recursos digitales con sumo cuidado.
- Evaluar el uso de estrategias de preservación tales como el rejuvenecimiento de los datos, verificaciones de consistencia de datos, la migración, emulación, preservación de la tecnología y arqueología digital.
- Considerar un enfoque híbrido.

Según Waugh y otros [10], las claves para la preservación de la información digital a largo plazo son:

- El encapsulado, es decir, empaquetar la información a ser preservada junto con metadatos descriptivos.
- Autodocumentación, es decir, la habilidad de entender y de codificar la información preservada sin referencia a documentación externa.
- Autosuficiencia, es decir, minimizar las dependencias de sistemas, datos o documentación.
- Documentación del tipo de contenido, es decir, la habilidad de un futuro usuario para encontrar o implementar software que permita ver la información preservada.

Según Helen Tibbo [11], la teoría de la archivística será esencial en el desarrollo de modelos de preservación intelectual a largo plazo de objetos digitales auténticos y confiables.

Cheney y otros [12] han incluso utilizado un lenguaje matemático para formalizar los conceptos que son relevantes a la preservación. Su teoría de los *espacios de preservación* se basa en ideas de la lógica y la semántica de los lenguajes de programación para describir la relación entre objetos concretos y su contenido de información. Se basan también en la teoría de los juegos para mostrar cómo los objetivos cambian con el paso del tiempo como resultado de efectos ambientales incontrolables y de acciones directas de preservación.

Cada autor tiene su enfoque y opinión personal sobre la preservación digital, pero casi todos coinciden en los mecanismos, los cuales veremos a continuación.

3.4 Preservación de los sistemas originales

El sistema más trivial para preservar los recursos electrónicos es mantener en funcionamiento el ordenador con el que han sido creados, almacenados, y pueden ser consultados. A pesar de ser una solución simple no es razonable querer mantener en funcionamiento a un equipo por décadas. A medida que pase el tiempo se volverá más difícil encontrar repuestos y las prestaciones que brinde el equipo se volverán obsoletas [10].

3.5 Emulación

Según Waugh y otros [10], la emulación permite que el software original sea usado sin necesidad de que el sistema original que lo ejecutaba siga existiendo. La emulación obliga a preservar una cantidad importante de información. Se debe preservar el emulador, el sistema operativo, la aplicación y los datos. No sólo es difícil identificar exactamente lo que debe ser preservado, sino que la pérdida de alguno de estos componentes hacen inaccesible la información. El emulador es también una aplicación de software, y deberá ser preservado, ya sea mediante emulación o mediante su actualización periódica. Esto puede convertirse en un cuento de nunca acabar, de emular la emulación de la emulación de la emulación...

3.6 Migración

La información digital es inútil, a menos que los formatos puedan ser reconocidos y procesados por un ordenador. Sabemos que los formatos de ordenador cambian continuamente y que algunos formatos y programas de hace más de diez años son difíciles de leer y ejecutar en la actualidad. La migración consiste en convertir la información a nuevos formatos. Es una medida contra la obsolescencia. Tiene la desventaja de ser una tarea pesada y de que los datos originales son modificados, con el peligro de que se produzcan efectos acumulativos no deseados tras múltiples migraciones.

3.7 Replicado y rejuvenecimiento

El replicado es una técnica básica de procesamiento de datos. Los datos importantes de los que existe sólo una copia en un ordenador son altamente vulnerables. El hardware puede fallar, los datos pueden ser dañados por software defectuoso o por un virus, por mala fe o por negligencia de un empleado, por alguna catástrofe o por simple envejecimiento del soporte físico. Por estas razones los centros de procesamiento de datos hacen rutinariamente copias de seguridad y las almacenan en lugares seguros. Debido a que todos los tipos de almacenamiento en los que se graba información digital son efímeros, las bibliotecas digitales deberían planear el rejuvenecimiento de sus colecciones periódicamente [13]. Cada pocos años los datos deben ser transferidos a nuevos medios de almacenamiento. Esta no es una exigencia excesiva desde el punto de vista económico dada la progresiva y continua bajada de precios y aumento de capacidad de los medios de almacenamiento, pero como suele ser el caso con las bibliotecas digitales, la duda es de organización: ¿serán las bibliotecas y las editoriales lo suficientemente sistemáticas y perseverantes como para llevar a cabo estos procesos?

3.8 Arqueología digital

Se llama arqueología digital al proceso de recuperar información a partir de fuentes de datos dañadas, fragmentadas o arcaicas. Es el remedio cuando no se han tomado los debidos recaudos y la información se ha estropeado.

3.9 Formatos digitales

La selección de formatos debe ser parte del plan global de preservación del proyecto, que debe abarcar también otros aspectos tales como los procedimientos de preservación a seguir. Entre estos, podemos citar, por ejemplo, la renovación periódica de los archivos para evitar la pérdida de datos debida al envejecimiento del soporte, la conversión de viejos formatos a otros más nuevos para evitar la obsolescencia, los criterios de redundancia, y la elección adecuada de medios y lugares de almacenamiento.

3.10 Una política adecuada

Arms [14] sugiere unos simples pasos para favorecer la longevidad de la información digital. El primero es almacenar la información en formatos que sean ampliamente usados hoy en día. Esto aumenta la probabilidad de que cuando un formato se vuelva obsoleto aún existan programas para su conversión. XML, HTML y PDF son ejemplos de estos. Otra interesante sugerencia es crear un archivo que contenga las definiciones de los formatos, estándares de metadatos, protocolos y otros elementos constructivos fundamentales de las bibliotecas digitales. Si los formatos y los esquemas de codificación son preservados, la mayoría de la información puede ser descifrada posteriormente.

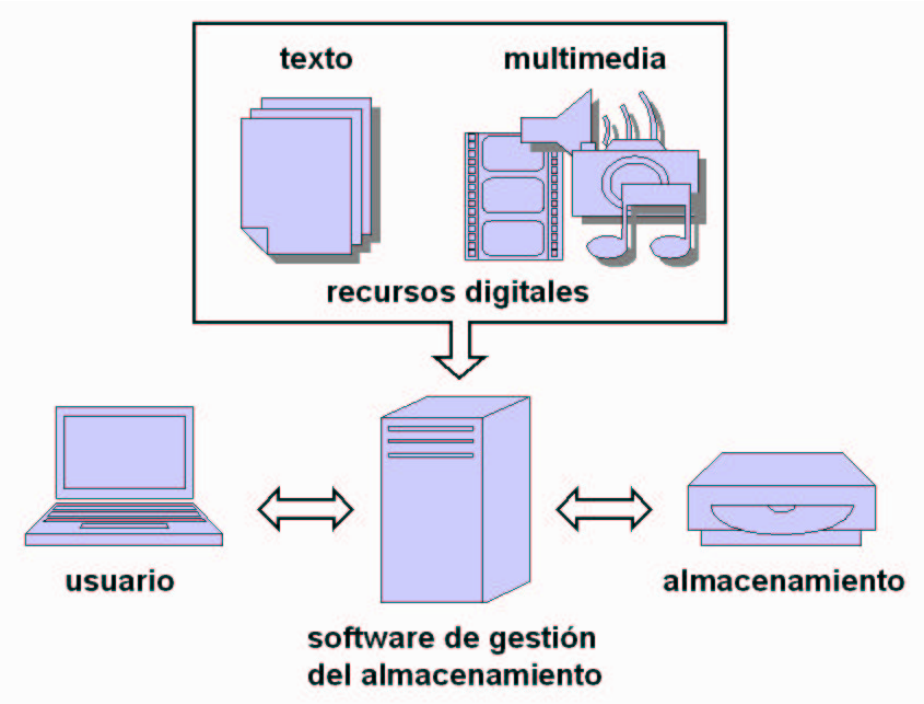

Fig. 1. Preservación digital

3.11 Acceso y recuperación de recursos almacenados

No tiene sentido tener algo almacenado si no podemos encontrarlo, o si ni siquiera sabemos que lo tenemos. El objetivo no es sólo preservar la información digital, sino tener un eficiente sistema de búsqueda y recuperación (ver figura 1).

4 Tipos de objetos digitales

La Biblioteca Virtual Miguel de Cervantes abarca una amplia diversidad de temas culturales y académicos que para su mejor aprovechamiento obligan al uso de una gran variedad de formatos digitales (multimedia) combinados de diferentes formas:

1. libros digitales en hipertexto[1] (XML --> HTML, PDF),
2. concordancias electrónicas[2],
3. facsímiles digitales[3],

[1] http://cervantesvirtual.com/catalogo.shtml
[2] http://cervantesvirtual.com/concordancias/index.shtml
[3] http://cervantesvirtual.com/bib_autor/Girondo/verfoto.formato?foto=graf/ilustraciones/ilus12.jpg

4. imágenes digitales (dibujos y fotos[4]),
5. imágenes y texto (p.ej. facsímiles con transcripciones[5]),
6. grabaciones de audio digital[6],
7. montajes de vídeo y texto sincronizados (biblioteca de signos[7]),
8. grabaciones de vídeo digital[8].

Cada tipo de objeto digital requiere un plan de preservación propio y adecuado.

5 Preservación de textos digitales

5.1 El uso de XML y la preservación de documentos

La elección de formatos debe ser parte del plan global de preservación del proyecto. Desde las primeras decisiones de diseño de nuestra biblioteca, uno de los aspectos importantes que hemos tenido en cuenta fue la elección de un formato de codificación de textos que favorezca la preservación de los mismos. Es necesario que el formato escogido para los documentos sea durable, y no un formato efímero fruto de una moda pasajera, que sea ampliamente usado y si es un estándar mejor, lo cual asegurará la existencia de programas que lo soporten, y que pueda fácilmente convertirse a los nuevos formatos que lo sucedan. Esto nos ha hecho excluir los formatos de propiedad de las empresas fabricantes de software[9] en favor de formatos abiertos de dominio público, ya que los primeros cambian según las caprichosas decisiones corporativas de los fabricantes y no son fáciles de convertir en otros formatos (cuando la opción de conversión no es proporcionada por el propio fabricante). Por ser formatos generalmente no documentados es difícil construir programas que los manejen. Por el contrario, el formato elegido debe ser durable y fácil de convertir a otros formatos. lo primero evitará que nos preocupemos de ello por un buen tiempo, y lo segundo asegurará que el día que sea necesario migrar a otro formato será fácil hacerlo. Estas condiciones pesaron al momento de elegir al XML como lenguaje de marcado para nuestros textos [15] (ver en la tabla 1 una comparación entre HTML, SGML y XML[10]).

Una solución fácil hubiera sido usar alguno de los editores comerciales más populares que permiten generar HTML y PDF para publicación web, pero éstos

[4] http://cervantesvirtual.com/bib_autor/Alfonsina/imagenes.shtml
[5] http://cervantesvirtual.com/bib_autor/Calderon/manuscrito/index.htm
[6] http://cervantesvirtual.com/bib_voces/bibvoces.shtml
[7] http://cervantesvirtual.com/portal/signos/
[8] http://cervantesvirtual.com/bib_imagenes/bibimagenes.shtml
[9] Usualmente llamados *proprietary formats* en inglés y mal llamados formatos propietarios en castellano.
[10] Existen varios artículos interesantes que comparan estos tres lenguajes de marcado de textos: Steven DeRose los compara de forma detallada en su artículo *XML and the TEI* [16], y James Clark presenta una comparación entre el SGML y el XML en su sitio web de la W3C [17]

Tabla 1. Comparación entre HTML, SGML y XML con respecto a los requerimientos de nuestra biblioteca digital (S=sí, N=no, E=se espera a corto plazo)

CARACTERÍSTICA	HTML	SGML	XML
Es fácil de usar	S	N	S
Es extensible	N	S	S
Se centra en la estructura	N	S	S
Es fácilmente convertible	S	S	S
No es un lenguaje propietario	S	S	S
Es ampliamente usado	S	N	E
Se espera que dure	S	S	S
Es directamente soportado por los navegadores	S	N	E

fueron descartados por razones de preservación entre otras. XML nos permite construir programas que generen otros formatos que los editores convencionales no proveen, como ficheros TACT para concordancias electrónicas. En cuanto a la tecnología de marcado, creemos que existe una tendencia a huir de la dependencia de los editores de texto convencionales a los que estábamos acostumbrados. Esta tendencia nos lleva a nuevos modos más controlables, transparentes y flexibles de codificar textos y objetos multimedia.

Según Morrison y otros [18], "El marcado es una parte crítica, e inevitable, de la creación y el procesamiento de textos. Independientemente del método de codificación de documentos elegido, alguna forma de marcado estará incluida en el texto. Si este marcado es *propietario* o *no-propietario*, basado en la apariencia o en el contenido depende de usted. Asegúrese de evaluar los objetivos del proyecto cuando tome las decisiones sobre codificación. Si el proyecto es de corto plazo o necesariamente dependiente del software, entonces las opciones son relativamente pocas. Sin embargo, si a usted le preocupa algo la preservación a largo-plazo, las capacidades multiplataforma, y/o el marcado descriptivo, entonces un lenguaje de marcado definible por el usuario (preferentemente TEI) es la mejor opción." Como Peter Shillingsburg corrobora: "... el editor que use un sistema de codificación universal para desarrollar una edición electrónica con una aplicación multiplataforma habrá creado una herramienta disponible para cualquier lector con acceso a un ordenador y habrá asegurado la longevidad de ese trabajo editorial para las generaciones de software y hardware venideras. Vale la pena el esfuerzo." [1].

En la Biblioteca Virtual Miguel de Cervantes, estamos convencidos de la importancia y el interés de preservar estos esfuerzos editoriales para las generaciones de lectores venideras.

References

1. Shillingsburg, P.: Scholarly Editing in the Computer Age: Theory and Practice. 3rd. edn. University of Michigan Press, Ann Arbor (1996)

2. Waters, D.J., ed. In: Digital archiving: the report of the CPA/RLG Task Force. National Preservation Office (1997)
3. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. 1st edn. ACM press and Addison Wesley, Edinburgh Gate, Harlow, Essex CM20 2JE, England (1999) See also http://www.dcc.ufmg.br/irbook or http://sunsite.dcc.uchile.cl/irbook.
4. Levy, D.M.: Heroic measures: reflections on the possibility and purpose of digital preservation. In: Proceedings of the third ACM conference on Digital libraries, Pittsburgh, Pennsylvania, United States, ACM Press (1998) 152–161
5. Rieger, O.Y.: 8. In: Projects to Programs: Developing a Digital Preservation Policy. Research Libraries Group, Mountain View, California, USA (2000) 150–152
6. Rieger, O.Y.: Libraries'Role in Preserving Digital Collections. Presentation at the Alicante meeting of the META-e Project (2001)
7. Kenney, A.R.: Digital to microfilm convertion: a demonstration project 1994-96. Cornell University Library (1996) http://www.library.cornell.edu/preservation/com/comfin.html.
8. Deegan, M., Tanner, S.: 8: Preservation. Digital Futures series. In: Digital Futures: strategies for the information age. Library Association Publishing, 7 Ridgmount Street, London WC1E 7AE (2002) 179–208
9. McCray, A.T., Gallagher, M.E.: Principles for Digital Library Development. Communications of the ACM **44** (2001) 49–54
10. Waugh, A., Wilkingson, R., Hill, B., Dell'oro, J.: Preserving Digital Information Forever. In: ACM 2000 Digital Libraries conference (Fifth ACM Conference on Digital Libraries), Menger Hotel, San Antonio, Texas, USA (2000) 175–184
11. Tibbo, H.R.: Archival Perspectives on the Emerging Digital Library. Communications of the ACM **44** (2001) 69–60
12. Cheney, J., Lagoze, C., Botticelli, P.: Towards a Theory of Information Preservation. In Constantopoulos, P., Solvberg, I., eds.: Research and Advanced Technology for Digital Libraries:5th European Conference, proceedings/ECDL 2001. Volume 2163 of Lecture Notes in Computer Science., Darmstadt, Germany, Springer-Verlag (2001) 340–351
13. Arms, W.: 13: Repositories and Archives. In: Digital Libraries. MIT Press, Cambridge, Massachusetts (2000) 245–262
14. Arms, W.: Digital Libraries. MIT Press, Cambridge, Massachusetts (2000)
15. Bia, A., Sánchez-Quero, M.: Diseño de un procedimiento de marcado para la automatización del procesamiento de textos digitales usando XML y TEI. In De-la-Fuente, P., Pérez, A., eds.: JBIDI 2001, II Jornadas Bibliotecas Digitales, Almagro (Ciudad Real), Spain (2001) 153–165
16. DeRose, S.: XML and the TEI. In Mylonas, E., Renear, A., eds.: Text Encoding Initiative: Anniversary conference; 10th — November 1997, Providence, RI. Volume 33(1) of Computers and the Humanities 1999; /2., Norwell, MA, USA, and Dordrecht, The Netherlands, Kluwer Academic Publishers Group (1999) 11–30
17. Clark, J.: Comparison of SGML and XML. http://www.w3.org/TR/NOTE-sgml-xml-971215 (1997)
18. Morrison, A., Popham, M., Wikander, K.: Oxford Text Archive, Creating and Documenting Electronic Texts. Oxbow Books, for the AHDS, Park End Place, Oxford OX1 1HN, UK (2000)