Procesos de digitalización y metadatos en el proyecto de la BDH

Ángel Ramos Arteaga, Pedro de Arce – Servicio de Biblioteca Digital
Biblioteca Nacional de España

ENCLAVE. Seminario sobre digitalización / Madrid / 10-11 Marzo 2010

Introducción

Las tareas de digitalización de las grandes colecciones se están realizando en el marco de programas europeos como "*2010: bibliotecas digitales*", con el objetivo de aumentar la digitalización de los fondos, y establecer estándares y motores de búsqueda que mejoren la accesibilidad y la preservación de la cultura europea.

En el caso de la Biblioteca Nacional, y gracias al patrocinio de Telefónica, en los próximos cuatro años se podrán consultar *online* las obras más importantes del patrimonio bibliográfico nacional: 15.000 manuscritos, 40.000 libros impresos de los siglos XVIII y XIX, 120.000 dibujos, grabados y fotografías, así como periódicos españoles e iberoamericanos, y se colgarán de la Red más de 25 millones de páginas.

Índice

01 Introducción
02 El punto de partida
03 Nuevos retos, nuevos flujos de trabajo
04 Pasos previos a la digitalización
05 Tratamiento de imágenes
06 Generación de metadatos
07 Modelo de exportación de metadatos. OAI-PMH
08 Sistema de gestión de objetos digitales
09 La BDH para el usuario final
10 Impacto exterior de la BDH

El punto de partida…

El extenso patrimonio de la Biblioteca Nacional

La colección de la Biblioteca se compone de más de 30.000 manuscritos, cerca de 3.000 incunables, unos 500.000 impresos anteriores a 1831, más de 6.000.000 de monografías modernas, cerca de 110.000 títulos de revistas y una colección de prensa estimada en casi 20.000 periódicos. La colección de partituras impresas y manuscritas supone más de 500.000 obras, los documentos sonoros en los diversos soportes depositados en el Biblioteca superan los 550.000 ejemplares y la colección de audiovisuales contiene más de 80.000 volúmenes, sin olvidar la colección de Bellas Artes y Cartografía.

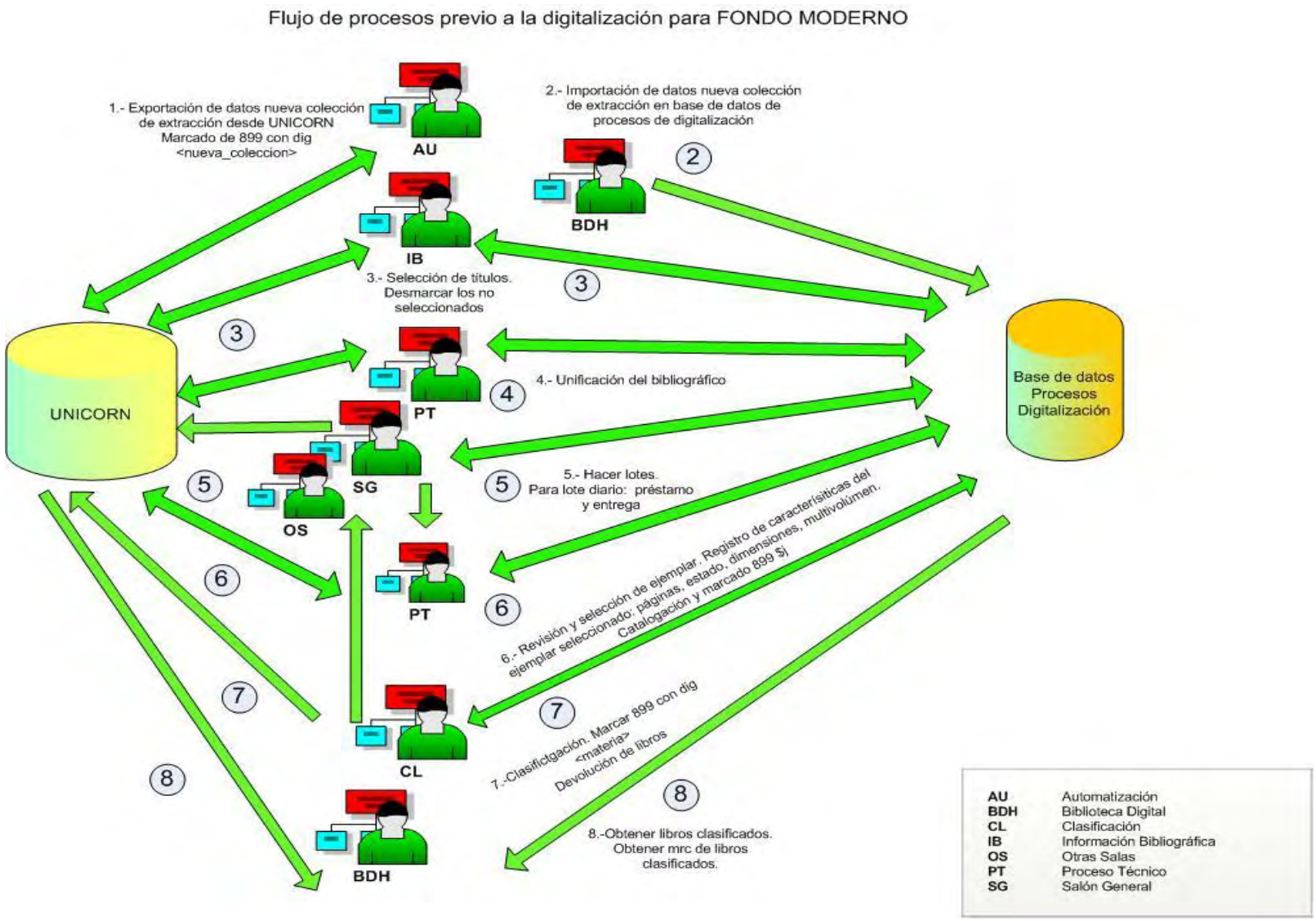
Manuscritos	Incunables	Impresos anteriores a 1831	Monografías modernas	Revistas	Periódicos
30.000	3.000	500.000	6.000.000	110.000	20.000

Nuevos retos, nuevos flujos de trabajo

En palabras de Milagros del Corral, Directora General de la Biblioteca Nacional, “digitalizar, no consiste sólo en que el libro o el grabado pase por un escáner. Hay que dotarlo de los oportunos metadatos de recuperación y de preservación del objeto digital", operación esta última que supone "una inversión brutal y un trabajo titánico" porque los formatos cambian constantemente.

La creación de colecciones digitales se enmarca dentro de un proceso transversal que, a grandes rasgos, puede resumirse en las siguientes etapas:

Procesos previos a la digitalización



Tratamiento de imágenes (1), preservación

TIPO DE DOCUMENTO	OBJETIVO	RESOLUCION	PROFUNDIDAD DE COLOR	NOTAS
Texto impreso SIN ilustraciones, prensa, panfletos, páginas mecanografiadas	Imagen del Texto	300 ppp mínimo	Escala de grises 8 bits *	*Color (24 bits) cuando el color sea una característica importante del documento
	Texto con OCR	400 ppi	Escala de grises 8 bits *	
Música: partituras, escalas anotadas, manuscritos de música	Acceso al contenido	300 ppp mínimo	Escala de grises 8 bits*	
	Reconocimiento de sus características materiales	400 ppi	Escala de grises 8 bits*	
Manuscritos: escritos a mano, copias mecanografiadas	Acceso al contenido	300 ppp mínimo	Escala de grises 8 bits*	*Color (24 bits) cuando el color sea una característica importante del documento
	Reconocimiento de sus características materiales	400 ppp	Escala de grises 8 bits*	
Mapas: caracteres impresos color impreso hasta un tamaño 56 cm x 87 cm	Búsqueda	250 ppp mínimo *	24-bit color	*La resolución (ppp) depende del tamaño del mapa, sobre todo en los casos en los que las secciones del mapa tienen que unirse y el tamaño del archivo sobrepase los 500 MB
	Reproducción	400 ppp	24-bit color mínimo	
Fotografías: tono continuo, color	Acceso al contenido	300 ppp mínimo	Escala de grises 8 bits*	*Color (24 bits) cuando el color sea una característica importante del documento
	Reproducción	Máximo soportado	24-bit color mínimo	
Material gráfico:	Acceso al contenido	300 ppi mínimo	Escala de grises 8 bits*	*Color (24 bits) cuando el color sea una característica importante del documento
	Reproducción	Máximo soportado	24-bit color	
Libros Especiales o Raros: Objetos de gran valor	Reconocimiento de sus características materiales	300 ppp mínimo	24-bit color	
	Investigación sobre sus características materiales	600 ppp mínimo	24-bit color mínimo	

Tratamiento de imágenes (2)

Tras el proceso de escaneado, el fichero TIFF-MASTER es recortado en dos partes, es decir, un fichero por cada página.

Pag. 1.

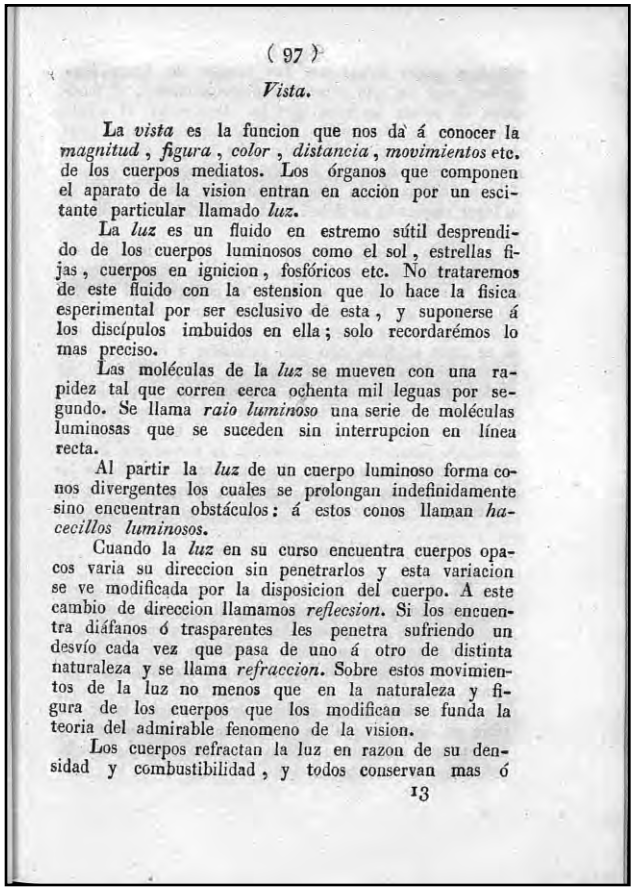
TRATADO

DE TRIGONOMETRIA.

INTRODUCCION.

LA TRIGONOMETRIA (que es una parte de la Geometria) es una ciencia, que enseña medir, y resolver triangulos. Y porque estos pueden ser planos, ò esphericos, se divide la Trigonometria en plana, y esphérica. Trigonometria plana es, la que enseña la resolucion de los triangulos planos, formados en una superficie plana con tres lineas rectas. Trigonometria esphérica es, la que dà reglas, para resolver los triangulos es-

A phe-



(97)

Vista.

La *vista* es la funcion que nos dà á conocer la *magnitud*, *figura*, *color*, *distancia*, *movimientos* etc. de los cuerpos mediatos. Los órganos que componen el aparato de la vision entran en accion por un escitante particular llamado *luz*.

La *luz* es un fluido en estremo sútil desprendido de los cuerpos luminosos como el sol, estrellas fijas, cuerpos en ignicion, fosfóricos etc. No trataremos de este fluido con la estension que lo hace la fisica esperimental por ser esclusivo de esta, y suponerse á los discípulos imbuidos en ella; solo recordarémos lo mas preciso.

Las moléculas de la *luz* se mueven con una rapidez tal que corren cerca ochenta mil leguas por segundo. Se llama *raio luminoso* una serie de moléculas luminosas que se suceden sin interrupcion en línea recta.

Al partir la *luz* de un cuerpo luminoso forma conos divergentes los cuales se prolongan indefinidamente sino encuentran obstáculos: á estos conos llaman *hacecillos luminosos*.

Cuando la *luz* en su curso encuentra cuerpos opacos varia su direccion sin penetrarlos y esta variacion se ve modificada por la disposicion del cuerpo. A este cambio de direccion llamamos *reflecsion*. Si los encuentra diáfanos ó trasparentes les penetra sufriendo un desvío cada vez que pasa de uno á otro de distinta naturaleza y se llama *refraccion*. Sobre estos movimientos de la luz no menos que en la naturaleza y figura de los cuerpos que los modifican se funda la teoria del admirable fenomeno de la vision.

Los cuerpos refractan la luz en razon de su densidad y combustibilidad, y todos conservan mas ó

13

Tratamiento de imágenes (3)

La aplicación de diferentes algoritmos permiten la mejora del texto sin pérdida de información.

Pag. 1.

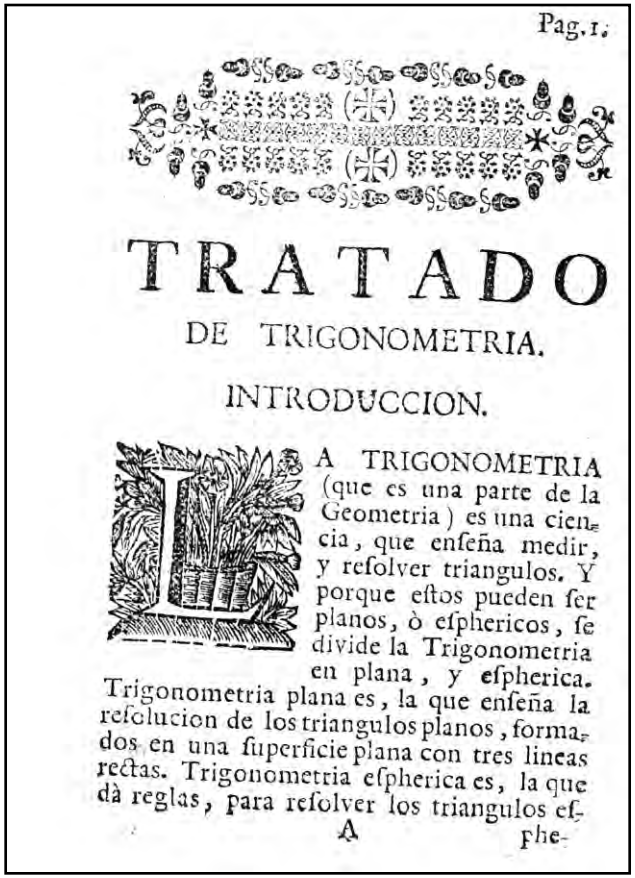
TRATADO

DE TRIGONOMETRIA.

INTRODUCCION.

LA TRIGONOMETRIA (que es una parte de la Geometria) es una ciencia, que enseña medir, y resolver triangulos. Y porque estos pueden ser planos, ò esphericos, se divide la Trigonometria en plana, y espherica. Trigonometria plana es, la que enseña la resolucion de los triangulos planos, formados en una superficie plana con tres lineas rectas. Trigonometria espherica es, la que dà reglas, para resolver los triangulos es-

A phe-



Pag. 1.

TRATADO

DE TRIGONOMETRIA.

INTRODUCCION.

LA TRIGONOMETRIA (que es una parte de la Geometria) es una ciencia, que enseña medir, y resolver triangulos. Y porque estos pueden ser planos, ò esphericos, se divide la Trigonometria en plana, y espherica. Trigonometria plana es, la que enseña la resolucion de los triangulos planos, formados en una superficie plana con tres lineas rectas. Trigonometria espherica es, la que dà reglas, para resolver los triangulos es-

A phe-

Tratamiento de imágenes (4)

Otros procesos llevados a cabo:

- Eliminación de manchas y suciedad
- Análisis y corrección de la inclinación de texto
- Extracción OCR
- Generación del fichero de difusión

Pag. 1.

TRATADO

DE TRIGONOMETRIA.

INTRODUCCION.

LA TRIGONOMETRIA (que es una parte de la Geometria) es una ciencia, que enseña medir, y resolver triangulos. Y porque estos pueden ser planos, ò esphericos, se divide la Trigonometria en plana, y espherica. Trigonometria plana es, la que enseña la resolucion de los triangulos planos, formados en una superficie plana con tres lineas rectas. Trigonometria espherica es, la que dà reglas, para resolver los triangulos es-

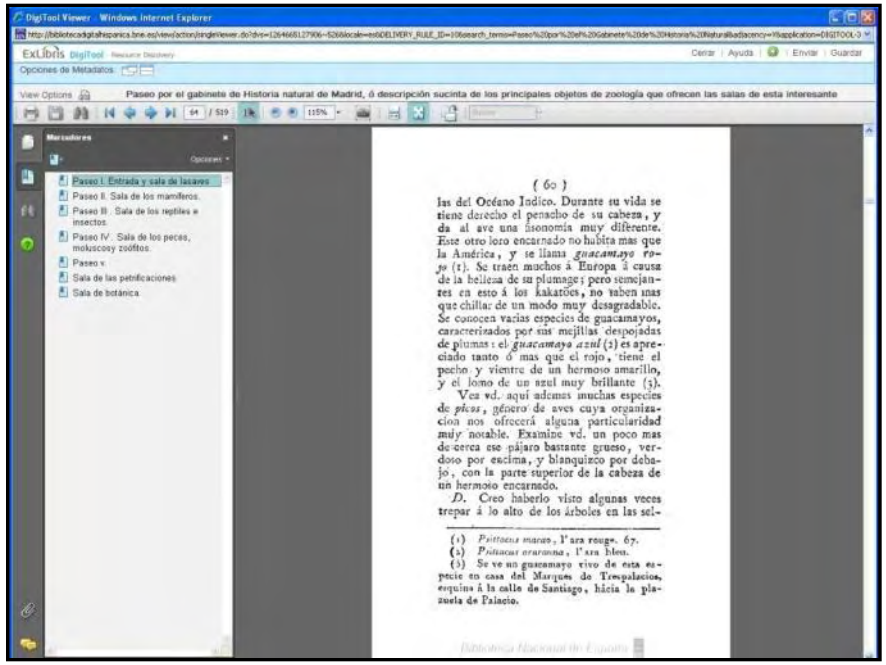
A phe-

Tratamiento de imágenes (5), difusión

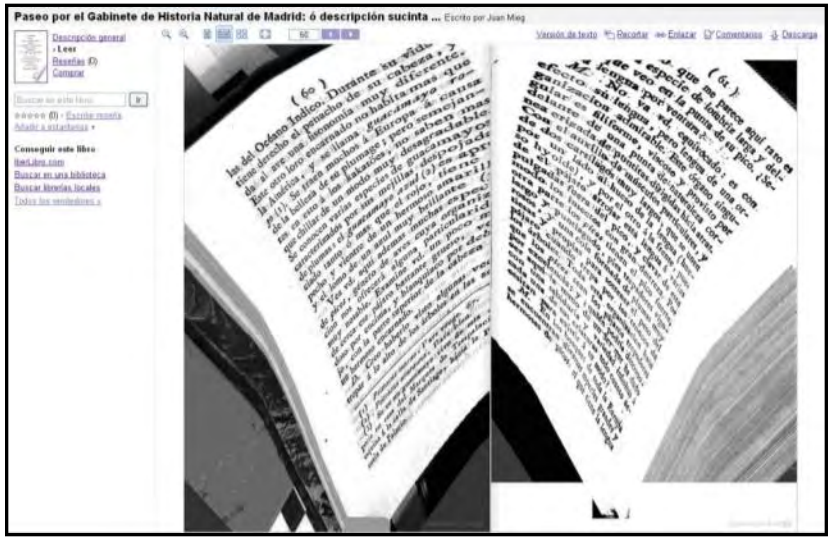
TIPO DE DOCUMENTO	ARCHIVO DE DIFUSIÓN
Texto impreso procedente de microforma	PDF con marcadores y OCR
Texto impreso procedente del original (incl. partituras impresas)	PDF con marcadores y OCR
Incunables digitalizados directamente del soporte original	PDF con marcadores sin OCR
Incunables digitalizados de microforma	PDF con marcadores sin OCR
Material gráfico digitalizado directamente del soporte original (grabados, estampas, dibujos, fotografías, carteles)	JPEG a 300 ppp
Material gráfico procedente de negativo	JPEG a 300 ppp
Mapas y planos	JPEG a 300 ppp En caso de que la toponimia y detalles del mapa o plano no se lean correctamente, se aumentará la calidad del JPEG
Manuscritos digitalizados directamente del soporte original (incl. música manuscrita)	JPEG a 300 ppp

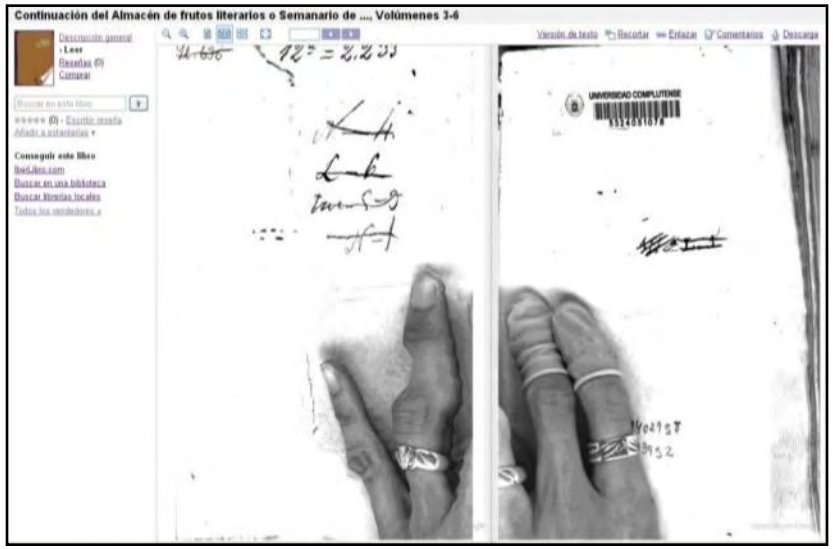
Tratamiento de imágenes (6), control de calidad

Biblioteca Digital Hispánica



Google Books





La BNE establece rigurosamente un importante control de calidad en todo el proceso destinado al tratamiento digital de las imágenes, tanto en la parte de preservación, como en la parte de difusión.

Generación de metadatos (1), metadatos descriptivos

1. A través de Unicorn, se generarán los metadatos descriptivos en formato mrc. (ISO 2709).

2. Estos ficheros mrc serán transformados a formato MARC21XML, con las especificaciones requeridas para su carga en el sistema de gestión de objetos digitales de la BNE.

3. Tras su correspondiente validación e integración con las imágenes correspondientes, se procederá a su ingesta en el sistema.

Generación de metadatos (2), metadatos DC

Dublin Core es un modelo de metadatos elaborado y auspiciado por la DCMI (Dublin Core Metadata Initiative), una organización dedicada a fomentar la adopción extensa de los estándares interoperables de los metadatos y a promover el desarrollo de los vocabularios especializados de metadatos para describir recursos para permitir sistemas más inteligentes del descubrimiento del recurso.

Dublin Core se define por ISO en su norma ISO 15836 del año 2003, y la norma NISO Z39.85-2007.

Es requisito imprescindible para formar parte de Europeana.

Generación de metadatos (3), PREMIS

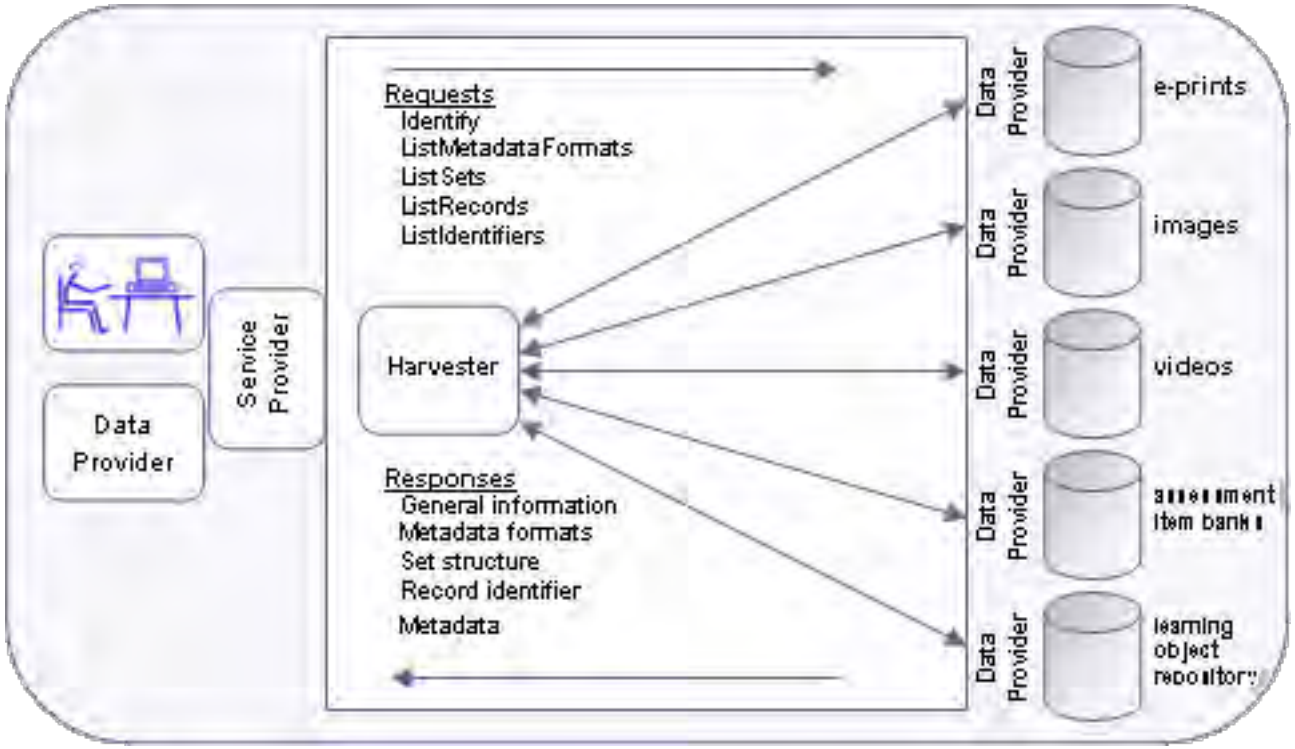
Los metadatos de preservación soportan las actividades cuyo objetivo es asegurar la utilización a largo plazo de un recurso digital.

PREMIS define los metadatos de preservación como ”la información que utiliza un repositorio para soportar el proceso de preservación digital”.

Ejemplo de PREMIS para la BDH


```
<?xml version="1.0" encoding="UTF-8" ?>
- <premis:premis version="2.0" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns:xlink="http://www.w3.org/1999/xlink" xmlns:premis="info:lc/xmlns/premis-v2"
xsi:schemaLocation="info:lc/xmlns/premis-v2 http://www.loc.gov/standards/premis/v2/premisv2-
0.xsd">
- <premis:object xsi:type="premis:representation" xmlID="VC_002307-006">
- <premis:objectIdentifier>
<premis:objectIdentifierType>899$j</premis:objectIdentifierType>
<premis:objectIdentifierValue>VC/2307/6</premis:objectIdentifierValue>
</premis:objectIdentifier>
- <premis:preservationLevel>
<premis:preservationLevelValue>full</premis:preservationLevelValue>
<premis:preservationLevelDateAssigned>20070529</premis:preservationLevelDateAssigned>
</premis:preservationLevel>
<premis:originalName>VC_002307-006</premis:originalName>
</premis:object>
- <premis:object xsi:type="premis:file">
- <premis:objectIdentifier>
<premis:objectIdentifierType>File</premis:objectIdentifierType>
<premis:objectIdentifierValue>VC_002307-006_0001</premis:objectIdentifierValue>
</premis:objectIdentifier>
- <premis:preservationLevel>
<premis:preservationLevelValue>full</premis:preservationLevelValue>
<premis:preservationLevelDateAssigned>20070529</premis:preservationLevelDateAssigned>
</premis:preservationLevel>
- <premis:objectCharacteristics>
<premis:compositionLevel>0</premis:compositionLevel>
<premis:size>1234567</premis:size>
- <premis:format>
- <premis:formatDesignation>
<premis:formatName>image/tiff</premis:formatName>
<premis:formatVersion>6.0</premis:formatVersion>
</premis:formatDesignation>
</premis:format>
- <premis:creatingApplication>
<premis:creatingApplicationName>Omniscan</premis:creatingApplicationName>
<premis:creatingApplicationVersion>11.0</premis:creatingApplicationVersion>
<premis:dateCreatedByApplication>20090102</premis:dateCreatedByApplication>
</premis:creatingApplication>
- <premis:objectCharacteristicsExtension>
- <mix:mix xmlns:mix="http://www.loc.gov/mix/v20"
xsi:schemaLocation="http://www.loc.gov/mix/v20
http://www.loc.gov/standards/mix/mix20/mix20.xsd">
- <mix:BasicDigitalObjectInformation>
<mix:byteOrder>big endian</mix:byteOrder>
- <mix:Compression>
<mix:compressionScheme>Uncompressed</mix:compressionScheme>
</mix:Compression>
</mix:BasicDigitalObjectInformation>
- <mix:BasicImageInformation>
- <mix:BasicImageCharacteristics>
<mix:imageWidth>5530</mix:imageWidth>
<mix:imageHeight>3210</mix:imageHeight>
- <mix:PhotometricInterpretation>
<mix:colorSpace>RGB</mix:colorSpace>
</mix:PhotometricInterpretation>
</mix:BasicImageCharacteristics>
</mix:BasicImageInformation>
- <mix:ImageCaptureMetadata>
- <mix:ScannerCapture>
<mix:scannerManufacturer>Zeutschel</mix:scannerManufacturer>
- <mix:ScannerModel>
<mix:scannerModelName>OS 10000-90 TT</mix:scannerModelName>
<mix:scannerModelSerialNo>52008</mix:scannerModelSerialNo>
</mix:ScannerModel>
</mix:ScannerCapture>
</mix:ImageCaptureMetadata>
- <mix:ImageAssessmentMetadata>
```

Modelo de exportación de metadatos. OAI-PMH

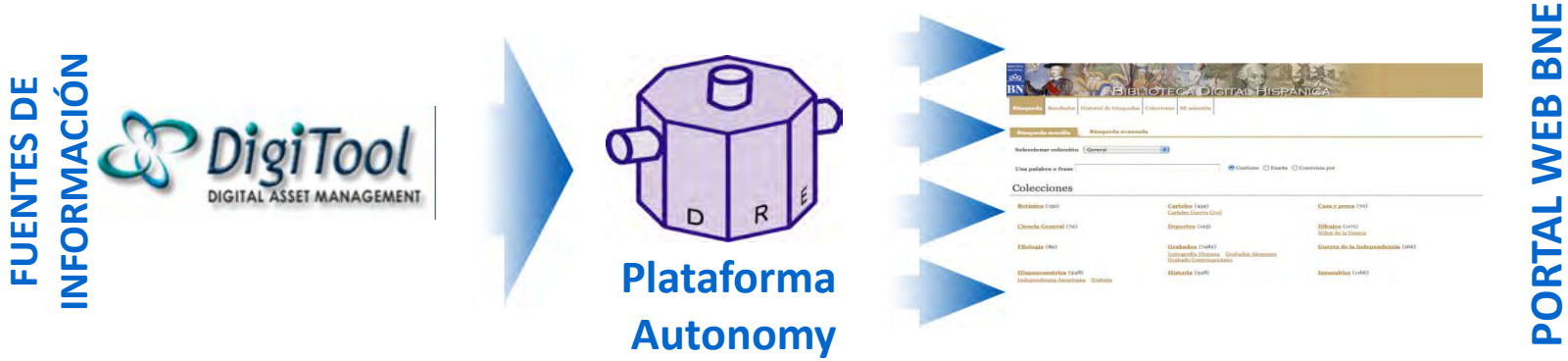


http://bibliotecadigitalhispanica.bne.es/OAI-PUB

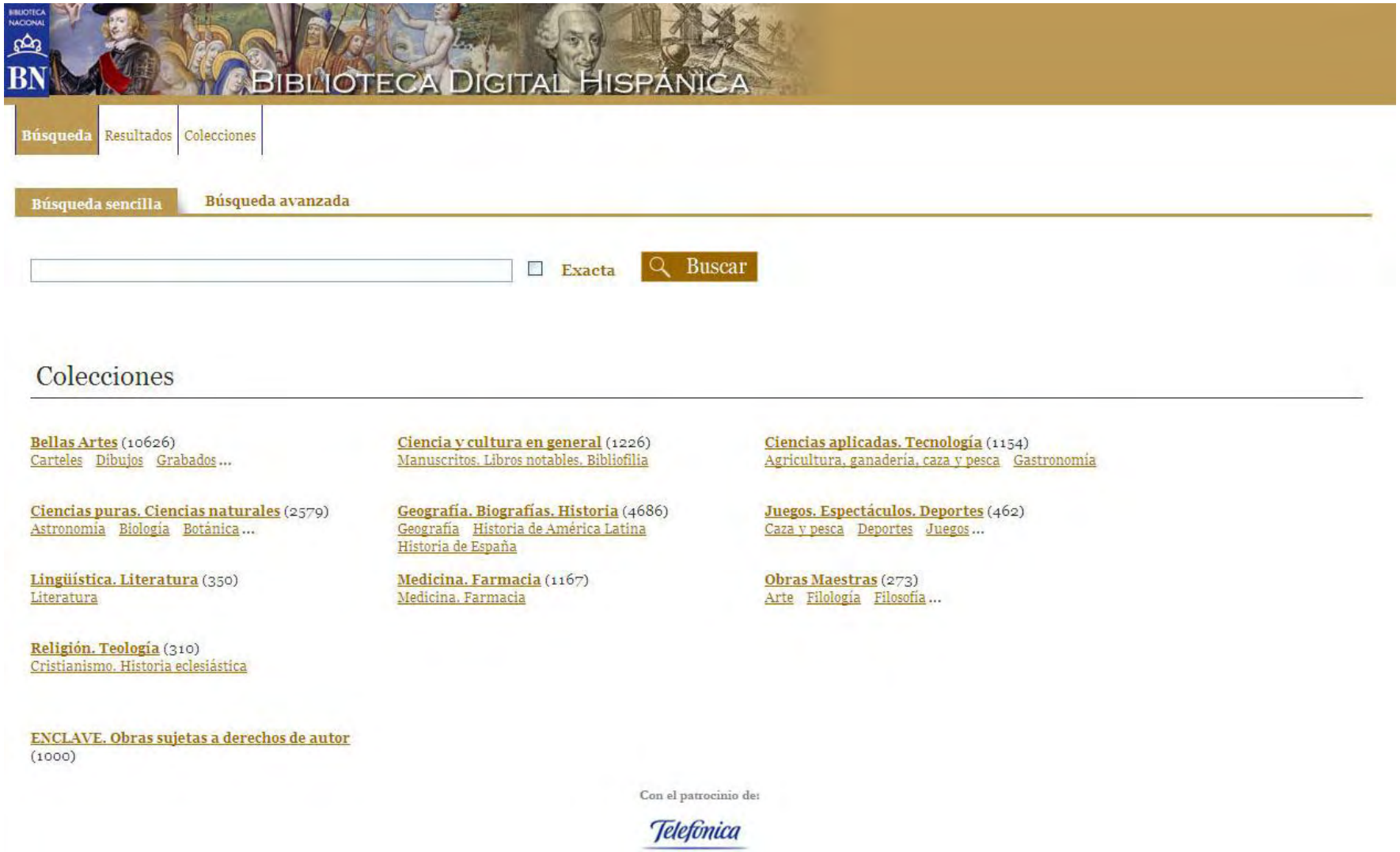
Sistema de gestión de objetos digitales

Características del Ingest:

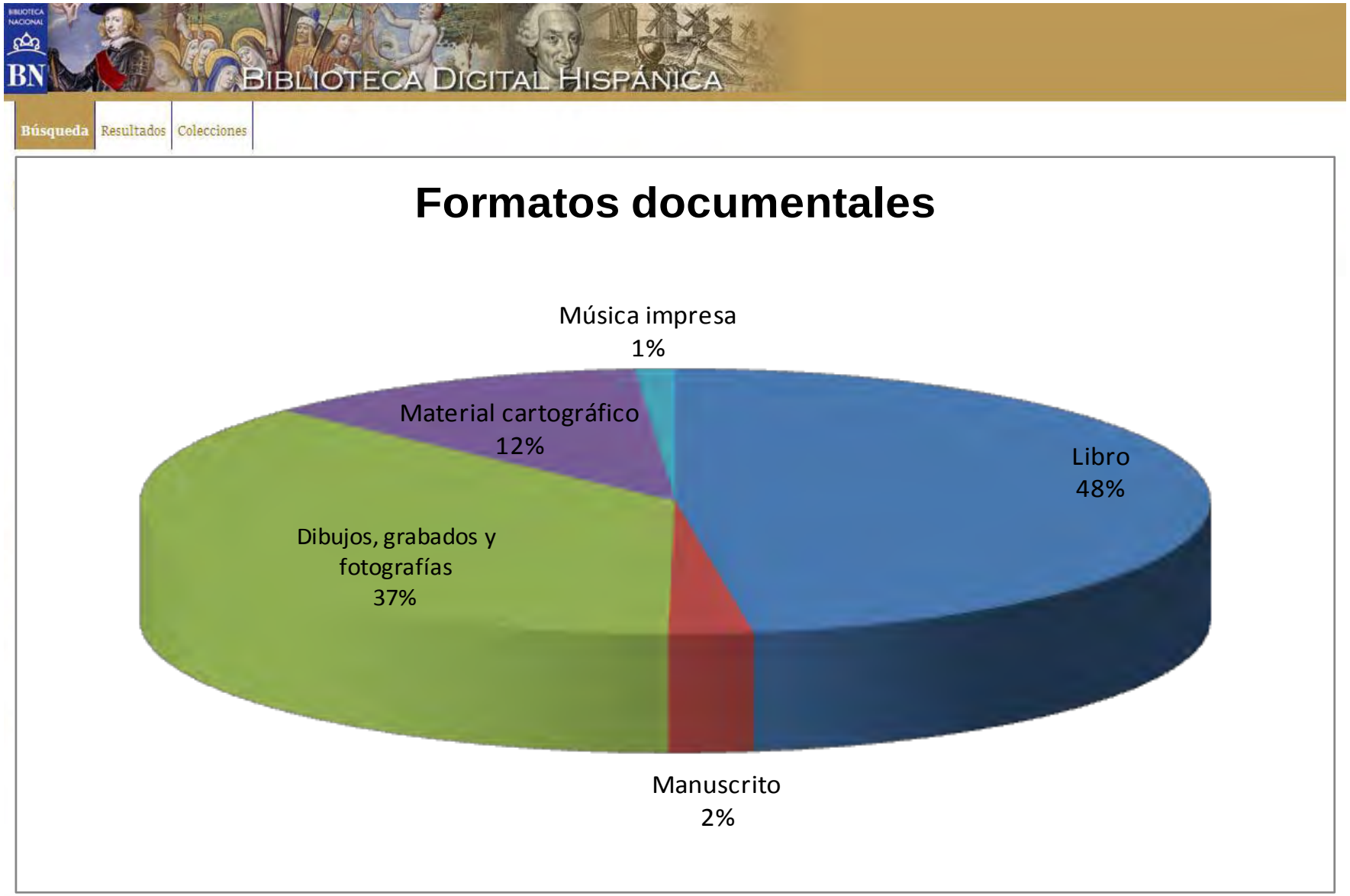
- Thumbnail Generation (Complex, Simples & METS)
- Add Metadata (Complex)
- Control section Attribute Assignment (Complex & Simples)
- Full Text Extraction (Complex, Simples & METS)
- Technical Metadata Extraction (Complex, Simples & METS)



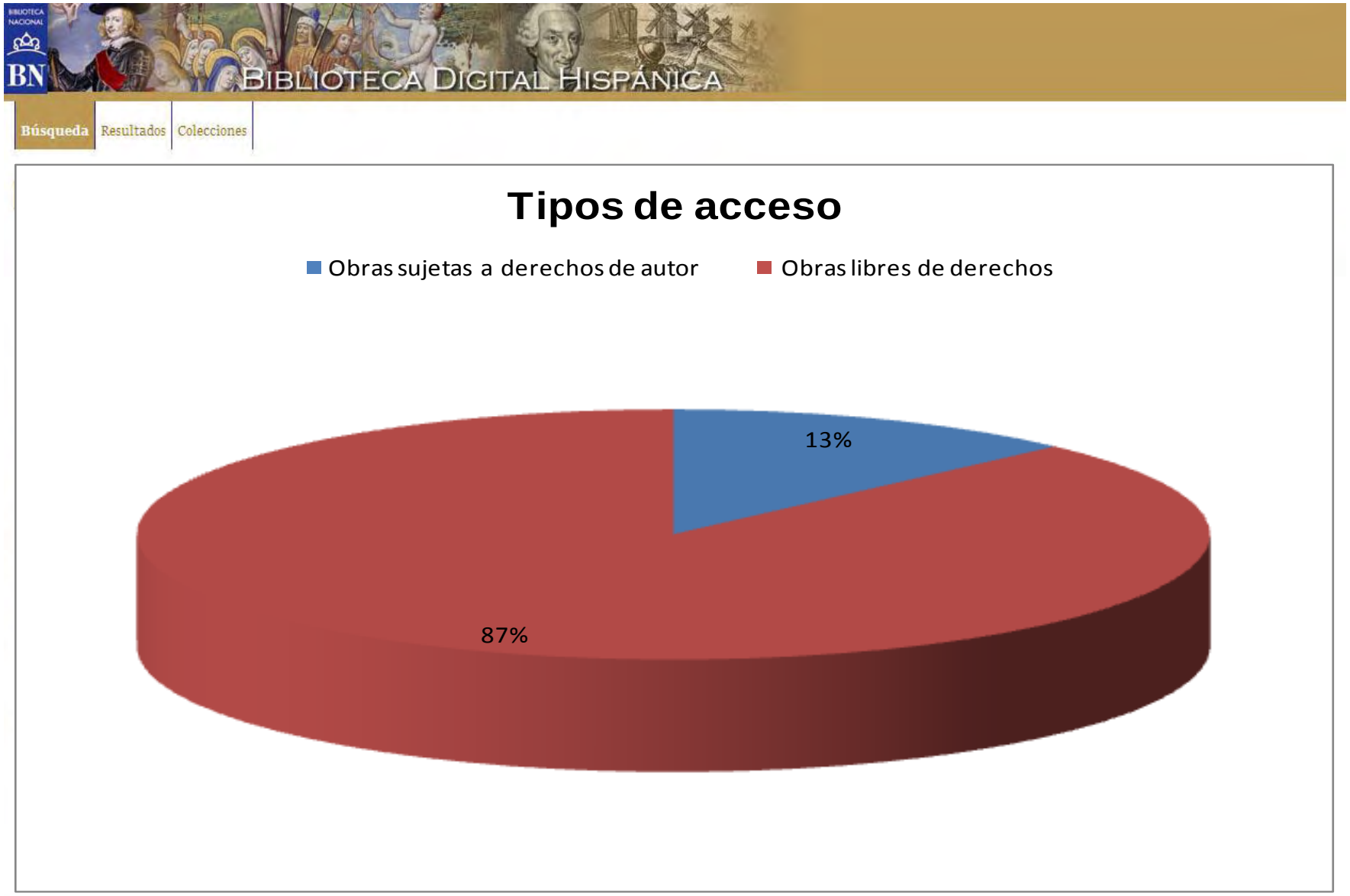
La BDH para el usuario final (1)



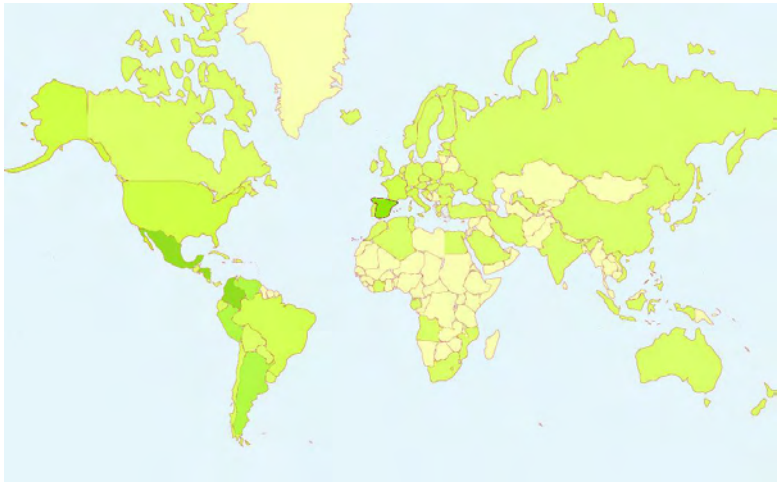
La BDH para el usuario final (2)

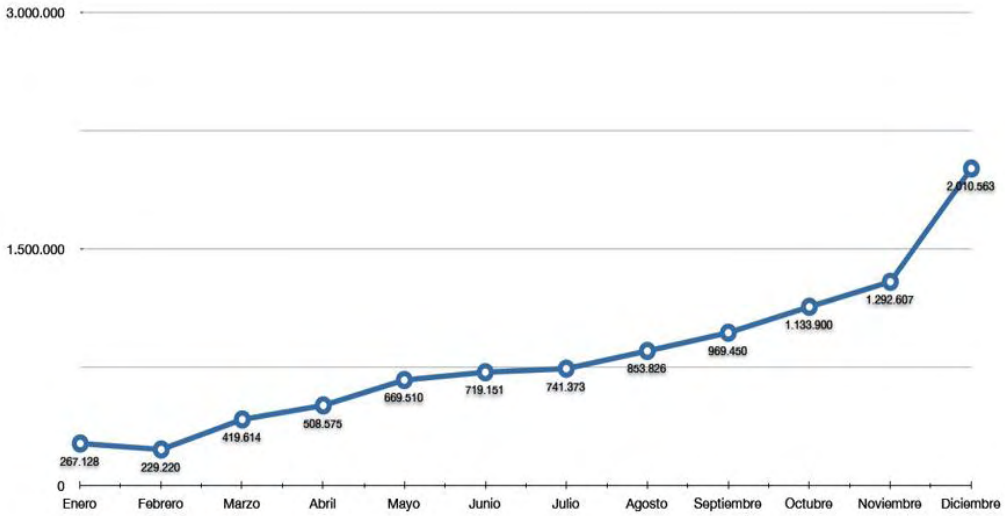


La BDH para el usuario final (3)



Impacto exterior de la BDH





Evolución mensual de accesos totales, Año 2009

Ángel Ramos Arteaga, Pedro de Arce Trujillo
Servicio de Biblioteca Digital
angel.ramos@bne.es
pedro.arce@bne.es

Pº de Recoletos 20 -22
28071 Madrid
España
T +34 915 807 800

www.bne.es