XUNTA DE GALICIA
CONSELLERÍA DE CULTURA, EDUCACIÓN
E ORDENACIÓN UNIVERSITARIA
Secretaría Xeral de Cultura

amtega
Axencia para a
Modernización Tecnolóxica

galicia

# REQUIRIMENTOS TÉCNICOS PARA PROXECTOS DE DIXITALIZACIÓN DE FONDOS BIBLIOGRÁFICOS EN INSTITUCIÓNS DA MEMORIA DE GALICIA

Xuño 2018

## ÍNDICE

## 1. INTRODUCIÓN

A Memoria Dixital de Galicia constitúese como un camiño global e colectivo de transformación, aproveitando a potencialidade das TIC, do modelo de xestión do patrimonio cultural e de dinamización da actividade cultural.

A construción dunha Memoria Dixital de Galicia é un proxecto colectivo de conservación, valorización e difusión do patrimonio cultural de Galicia. Este proxecto non se centra unicamente na fase de dixitalización, senón que ten en conta a totalidade do proceso do patrimonio cultural dixital, asumindo unha necesaria coordinación e complementariedade entre a dixitalización, a preservación dixital e difusión dos activos resultantes. A Memoria Dixital de Galicia nace coa vontade de ser un proxecto con continuidade no tempo, transformando a relación cos usuarios e favorecendo a implicación da cidadanía.

A Memoria Dixital de Galicia é unha estratexia impulsada desde o ámbito público, a través da Consellería de Cultura, Educación Ordenación Universitaria e a Axencia para a Modernización Tecnolóxica de Galicia (Amtega) que busca a colaboración coa sociedade galega e con todos os axentes implicados na xestión do patrimonio cultural galego.

Para construír a Memoria Dixital de Galicia como novo modelo de xestión do patrimonio cultural galego baseado nas TIC identificáronse 5 liñas de actuación:

1º - Definir un marco técnico común para todos os axentes e institucións involucrados na conservación e difusión do patrimonio cultural mediante o establecemento de políticas, normas, procedementos e estándares para a recolleita, catalogación, preservación e difusión do patrimonio cultural galego.

2º - Aplicar as novas tecnoloxías para a catalogación, inventariado e xestión do patrimonio, o que implica a dotación de infraestruturas e ferramentas tecnolóxicas nas institucións da memoria que custodian os bens patrimoniais, así como a capacitación no seu uso.

3º - Aproveitar as TIC para facilitar a visibilización e a promoción do patrimonio cultural, fundamentalmente mediante o impulso de políticas de dixitalización do patrimonio e tamén fomentando a aplicación ou desenvolvemento de técnicas novas e interactivas para o acceso, conservación, restauración e posta en valor do patrimonio cultural.

4º - Aplicar a tecnoloxía para mellorar a accesibilidade ao patrimonio desde o punto de vista presencial e virtual. No primeiro caso, para mellorar os servizos prestados directamente en arquivos, bibliotecas, museos e calquera outra institución que custodie os recursos patrimoniais. No segundo, incrementando e mellorando as solucións dixitais de acceso ao patrimonio, adaptándoas ao perfil e ás necesidades do

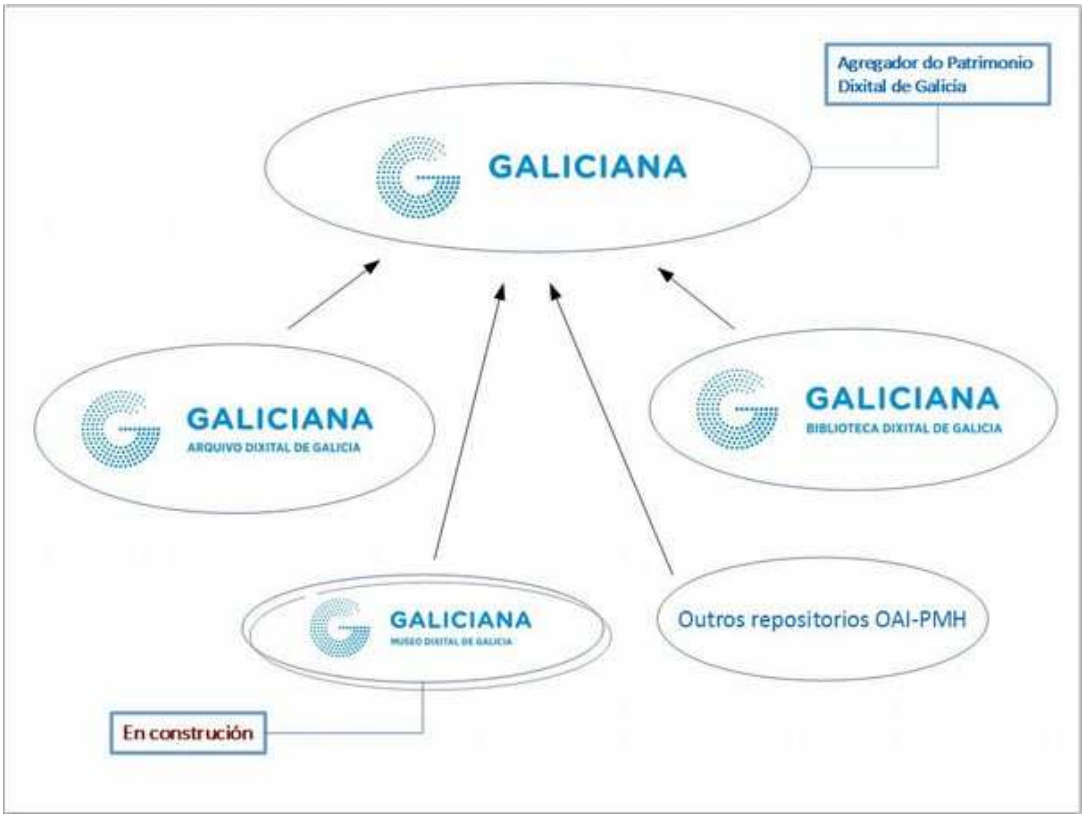
usuario, desde o público en xeral a profesionais, docentes, investigadores...

5º - Fomentar os uso das TIC como elemento dinamizador da produción cultural, impulsando a creación de contidos dixitais culturais e a súa correspondente promoción, e que contribuirá a xerar novas oportunidades de desenvolvemento económico na sociedade galega.

Tal e como se establece na 1º liña de actuación, un dos obxectivos que pretende acadar a Memoria Dixital de Galicia é o *"establecemento de políticas, normas, procedementos e estándares para a recolleita, catalogación, preservación e difusión do patrimonio cultural galego"*. Este documento trata de cumprir con este obxectivo marcándose como finalidade:

- Normalizar as características técnicas da dixitalización e os estándares mínimos de asignación de metadatos para así facilitar o intercambio de recursos e facer posible as buscas a través dunha única plataforma (Galiciana-Patrimonio Dixital de Galicia).
- Asesorar e coordinar os esforzos que as diferentes institucións da memoria de Galicia estean a levar adiante en materia de dixitalización, de forma que se optimicen os recursos persoais, económicos e materiais.

## Estrutura de Galiciana-Patrimonio Dixital de Galicia



É importante sinalar que este documento céntrase nas pautas a seguir para unha correcta dixitalización e identificación mediante esquemas de metadatos dos **materiais vinculados ao eido bibliotecario** (monografías, folletos, mapas, postais, publicacións periódicas, carteis ou fotografías).

Estas recomendacións non pretenden ser un manual de dixitalización senón simplemente un instrumento para o traballo en común e a coordinación dos proxectos que se están a realizar ou se realicen no futuro dentro do ámbito bibliotecario galego. Trátase dun documento adaptado ás características e panorama actual das normativas e directrices nacionais e internacionais vixentes para estes procesos e que requirirá de posteriores actualizacións que permitan adaptar o seu contido a futuros desenvolvementos neste ámbito de aplicación.

A existencia de pautas e directrices xa publicadas e adoptadas tanto a nivel nacional como internacional, aconsella non establecer novas normativas que poidan diverxer dos ámbitos comúns xa establecidos. En consecuencia, este documento segue e respecta esas directrices xa consolidadas e extrae as súas principais ideas das que se citan na bibliografía.

De xeito particular, para a redacción destes requisitos foi fundamental o establecido nos documentos elaborados pola Subdirección General de Coordinación Bibliotecaria, *Requisitos técnicos para proyectos de digitalización de patrimonio bibliográfico y de prensa histórica de la SGCB (v. 1.1.2. de 11.05.2016),* e pola Biblioteca Nacional de España, *Proceso de digitalización en la Biblioteca Nacional de España (actualizado a 25.02.2015).*

En resumo, a función desta publicación está, por un lado, en divulgar no noso ámbito unha serie de conceptos que teñen unha importancia básica para o deseño dunha política común de dixitalización do patrimonio cultural e por outro, concretar ou desenvolver aqueles aspectos desta política que poidan axudar a levar adiante procesos de dixitalización do patrimonio bibliográfico e documental na nosa comunidade.

## 2. DEFINICIÓNS

**Dixitalización**: proceso de converter unha sinal analóxica en dixital.

**Imaxe dixital**: foto electrónica tomada dunha escena ou escaneada dalgún documento (fotografías, manuscritos, textos impresos, ilustracións).

**Captura da imaxe**: proceso polo que se obtén unha representación dixital dun orixinal constituída por un conxunto de elementos pictóricos ou píxeles mediante o escaneado ou fotografía dixital.

**Píxel** (abreviatura de "picture element"): Elemento mínimo dos que, en conxunto, forman unha imaxe dixital.

**Ppp** (puntos por polgada) **= Dpi** (dots per inch,): medida da resolución espacial das imaxes. Sirve para medir a resolución que é a cantidade de puntos (píxeles) que entran nunha polgada.

**Resolución:** indica o número de píxeles utilizados para representar a imaxe, medidos en proporción á unha superficie determinada (píxeles por polgada). A resolución é a capacidade dunha imaxe de representar o detalle. Cantos máis píxeles ou puntos se utilicen para definir unha polgada da imaxe, máis resolución terá esta. En xeral, a máis resolución, máis fiel ao orixinal será a imaxe dixital. A maior resolución a imaxe terá máis peso e tamén máis custo de produción, polo que é importante considerar para cada material a resolución máis adecuada.

**Compresión:** permite reducir o tamaño do ficheiro da imaxe para o seu almacenamento e transmisión. A compresión pode ser con perda ou sen perda. Sempre se recomenda empregar sistemas de compresión estandarizados.

**Compresión con perda:** proceso que reduce o espazo de almacenamento necesario para o ficheiro dunha imaxe mediante a eliminación de datos desa imaxe. Ao descomprimir unha imaxe que experimentase unha compresión con perda sempre será distinta da imaxe antes de comprimila, incluso anque a diferencia sexa difícil de detectar para o ollo humano.

**Compresión sen perda:** proceso que reduce o espazo de almacenamento necesario para o ficheiro dunha imaxe sen perda de datos. Se unha imaxe experimentou unha compresión sen perda, será idéntica á imaxe antes de comprimirse.

**Ficheiros máster:** o ficheiro máster é a versión da imaxe dixital de alta calidade sen compresión. A súa función é a de servir como copia de preservación e representar coa máxima fidelidade posible o orixinal. A partir destes ficheiros xeneraranse os outros ficheiros de imaxe de menor calidade.

**Ficheiros derivados:** obtéñense dos ficheiros máster e son os que se empregan para a consulta e difusión a través das redes. O requisito fundamental destes ficheiros é que son máis lixeiros despois de pasar por un proceso de compresión con perda.

**Ficheiros de miniaturas:** ficheiros dixitais que representan as imaxes en versións pequenas, de baixa resolución.

**Bits:** unidade mínima de información dixital. Un bit só pode tomar os valores 0 e 1.

**Profundidade de cor:** é o número de bits utilizados para representar a cor de cada píxel. A máis profundidade, máis detalles cromáticos.

Bitonal: 1 bit (branco ou negro)
Escala de grises: entre 2 e 8 bits (ata 256 grises)
Cor: 24 bits (16.7 millóns de cores)

O mesmo que coa resolución, cada tipo de material require unha definición óptima que permita explotar ó máximo as súas posibilidades de información é o tempo non desaproveite recursos innecesarios.

**Tipo de cor:** branco e negro, escala de grises ou color, en relación tamén coa profundidade escollida.

**Formato:** os datos recollidos polo software do escáner, deben ser almacenados de acordo cun formato. Os formatos de arquivo son unidades lóxicas de almacenamento da información que consisten tanto nos bits que comprende a imaxe como os datos acerca de como ler e interpretar o arquivo. Varían en termos de resolución, profundidade de bits, soporte para compresión e metadatos. A elección do formato dependerá da finalidade da imaxe dixital (preservación, difusión,...). Hai que escoller formatos estándar. Considérase unha boa práctica que na fase de captura o formato de saída sexa de alta calidade e sen compresión, do cal poidan derivarse logo os subprodutos (imaxes para a difusión, miniaturas e PDF).

**TIFF** (Tagged Image File Format): formato de ficheiros para imaxes con etiquetas. Isto se debe a que os ficheiros TIFF conteñen, ademais dos datos da imaxe propiamente dita, "etiquetas" nas que se arquiva información sobre as características da imaxe, que sirve para o seu tratamento posterior. Este formato é de aplicación xeneralizada para a creación de imaxes de alta calidade, produce ficheiros de gran tamaño, sen perda, útiles como ficheiros máster pero inadecuados para a distribución e acceso público ás coleccións. Pode presentar calquera resolución, branco e negro, escala de grises ou cor.

**JPEG** (Joint Photographic Experts Group): norma para a compresión da imaxe con calidade fotográfica na World Wide Web. É un formato para o almacenamento e transmisión de imaxes na Web. O seu algoritmo de comprensión permite reducir o tamaño dos ficheiros, sen perda ou cunha perda pouco significativa da calidade da imaxe.

**PDF** (Portable Document Format): é un formato de almacenamento de documentos desenvolvido pola empresa Adobe Systems, especialmente adecuado para a presentación de documentos complexos (múltiples páxinas, combinación de textos e imaxes de diferentes calidades). Este formato ofrece, entre outras vantaxes, opcións de navegación no documento e entre diferentes documentos, fidelidade e seguridade da copia dixital e posibilidades de busca e recuperación a partir dos contidos.

**Metadatos:** conxunto de informacións relacionadas cos obxectos dixitais. Representan ao obxecto dixital con fins de descrición, xestión e intercambio. Divídense en diferentes grupos segundo a finalidade que sustentan (procura e recuperación da información, xestión de obxectos dixitais, estruturación física e lóxica dos mesmos). O seu obxectivo e garantir a interoperabilidade, visibilidade, autenticidade e preservación dos obxectos dixitais.

**Metadatos descritivos:** esquemas especialmente destinados a procura e recuperación da información e á recolleita dos datos. Son os metadatos que describen o contido físico e conceptual dun obxecto. Algúns dos esquemas de metadatos descritivos máis empregados son MARCXML, DC, MODS ou ONIX. Poden utilizarse por si mesmos ou formando parte doutros esquemas e protocolos.

**Metadatos administrativos:** son aqueles que describen a procedencia dun obxecto dixital, os procesos realizados para a súa creación ou xeración, as súas características técnicas, as súas condicións de acceso e dereitos de propiedade intelectual e as accións xa realizadas ou previstas relacionadas coa preservación do obxecto en si mesmo. Comprenden tanto os metadatos técnicos como os de propiedade intelectual e de preservación. Xeralmente empréganse para a xestión interna dos ficheiros dixitais.

**Metadatos técnicos:** describen os atributos do obxecto dixital. Son os metadatos que proporcionan a información sobre o proceso de captura (data, axentes que dixitalizaron, ...) e os elementos técnicos (hardware e software) empregados para escanear os documentos orixinais. Tamén facilitan datos sobre os formatos e a calidade das imaxes.

**Metadatos de preservación:** datos que permiten seguir o ciclo de vida dos obxectos dixitais para garantir a súa accesibilidade futura, a súa correcta interpretación e a súa autenticidade e integridade, calquera que sexa o sistema ou metodoloxía que se empregue para elo. Conteñen a información que utiliza un repositorio para soportar o proceso de preservación. Diferéncianse dos metadatos técnicos no sentido de que documentan tamén os procesos que sofren os obxectos dixitais ao longo do tempo (copias, migracións ou calquera outra alteración), incluso despois de importados a un repositorio.

**Metadatos estruturais:** son os que permiten describir as relacións entre as diferentes partes dun recurso dixital. Achegan información sobre a estrutura intelectual dun obxecto (páxinas, capítulos, ...) e tamén permiten asociar os diferentes formatos que poidan existir dese obxecto.

**METS** (Metadata Encoding and Transmission Standard): o esquema METS é un estándar para a codificación dos metadatos descritivos, administrativos e estruturais dos obxectos dunha biblioteca dixital. Para elo emprega XML como linguaxe de marcado. Xurde no 2002 como unha iniciativa da *Digital Library Federation* para proporcionar os datos necesarios, nun contorno XML, para a xestión dos materiais dixitais. Permiten tanto o almacenamento dos datos dos obxectos dixitais, como o intercambio de información entre bases de datos ou a difusión aos usuarios finais. É un estándar que mantén a *Network Development and MARC Standards Office* da *Library of Congress*.

**MARC 21 XML (MARC XML):** esquema que permite reproducir un rexistro MARC completo en XML.

**MODS (Metadata Object Description Schema):** esquema de metadatos descritivos que se deriva do MARC 21. Os seus elementos teñen unha presentación máis lexible que os do formato MARC XML (<titleinfo> en lugar de <datafiled tag=”245” ind1=”1” ind2=”0”). A súa principal vantaxe é a súa capacidade para establecer relación xerárquicas entre as descricións e proporcionar unha codificación detallada das partes dun recurso (volume, número, capítulo, ...).

**PREMIS (Preservation Metadata Implementation Strategies):** metadatos de preservación que conteñen a información que utiliza un repositorio para xestionar o proceso de preservación dixital. Están baseados no modelo de referencia OAIS (Open Archival Information System), o cal establece a necesidade de rexistrar unha serie de datos mínimos que permitan seguir o ciclo de vida dos obxectos dixitais para garantir a súa accesibilidade, interpretación, autenticidade e integridade futuras.

**OAI-PMH (Open Archives Initiative–Protocol Metadata Harvesting):** é un protocolo de recolección sinxelo para o intercambio de metadatos entre repositorios. Os metadatos a recolectar poder estar en calquera formato establecido por calquera conxunto específico de provedores de datos e provedores de servizos), con independencia de que, como mínimo, deben codificarse en Dublin Core no cualificado para proporcionar un nivel básico de interoperabilidade.

**OCR (Optical Character Recognition–Recoñecemento Óptico de Caracteres):** tecnoloxía que permite ler os caracteres de texto individuais dunha páxina en formato dixital e converter esa información nun ficheiro de texto que pode ser almacenado de forma electrónica.

**ALTO (Analyzed Layout and Text Object) XML:** esquema para especificar os metadatos técnicos do deseño e contido dos recursos textuais físicos, tales como as páxinas dun libro ou dun periódico, e que permiten codificar en XML o resultado dun proceso de OCR.

# 3. DIXITALIZACIÓN

A dixitalización é un proceso no que inciden tanto elementos de carácter informático, referentes especialmente ás características técnicas do proceso e resultados da dixitalización en si, como documentais que afectan aos metadatos que achegamos ás imaxes para facilitar a súa descrición e posterior recuperación polos usuarios das bibliotecas dixitais.

Neste punto trataremos de aclarar aspectos concretos do proceso de obtención das imaxes dixitais mentres que o seguinte centrarase nos elementos documentais da dixitalización.

Contar con perfís técnicos comúns para os proxectos de dixitalización que se desenvolvan en Galicia é imprescindible para que todos poidan estar presentes en Galiciana, ventá de acceso en Internet ao patrimonio dixital de Galicia.

## 3.1. Obtención das imaxes

- Deberán empregarse escáneres aéreos e de tamaño axustado aos orixinais.
- Xerarase unha copia das obras dixitalizadas en formato TIFF 6.0/ISO 12639:2004 e outra en formato JPEG/ISO/IEC 10918:1994 como formatos de máxima calidade e formato comprimido e optimizado para a súa lectura en Internet respectivamente. Tamén se proporcionarán miniaturas de cada unha das imaxes dixitalizadas e un arquivo en formato PDF de cada unha das obras.
- As imaxes das obras non se entregarán a dobre páxina. No caso de que o fondo seleccionado permita a súa dixitalización a dobre páxina (manuscritos, monografías e incunables), do proceso de escaneado obterase un ficheiro TIFF máster e un TIFF recortado en dúas partes, é dicir, un ficheiro por cada páxina.
- De ser necesario, realizarase o enderezamento das imaxes para lograr que a copia dixital sexa o máis fidedigna posible ao orixinal.
- Recoméndase o uso de elementos de control, como cartas de grises ou de cores ou escalas métricas, que se inclúen en procesos de dixitalización de materiais textuais o gráficos para conseguir unha representación fiel do orixinal.
- As páxinas que se dixitalicen en formato TIFF xunto a estes elementos de control, dixitalizaranse dúas veces, unha cos referidos elementos e outra sen eles.
- A imaxe TIFF na que aparezan as cartas de grises ou de cores ou escalas métricas irá tamén apropiadamente referenciada no ficheiro METS de preservación asociado ao correspondente obxecto dixital.
- Nos formatos derivados dos distintos obxectos dixitais (JPEG, PDF e MINIATURAS) non será necesario incluír a imaxe cos elementos de control empregados para garantir a máxima calidade na dixitalización.
- Recortaranse os bordes negros das imaxes, pero non de xeito automático, xa que deberá ofrecerse ao usuario unha imaxe o máis aproximada posible á forma das páxinas orixinais.
- Asegurarase a correcta manipulación dos exemplares. En ningún caso poderán ser

desencadernados ou guillotinados e evitarase o emprego de prensalibros.

- No proceso de dixitalización das obras usarase unha cartolina de cor negro debaixo dos exemplares para visualizar de forma correcta os bordes das obras. No caso de que o soporte estivese incompleto ou deteriorado interporanse follas de papel xaponés para evitar ver as seguintes páxinas.
- Os ficheiros cas marcas de auga incluiranse nas respectivas imaxes JPEG para difusión en web, nunca nas copias máster. En todo caso, serán de pequeno tamaño e non se superpoñerán ao contido do documento.
- Os requisitos mínimos para a obtención das imaxes segundo o tipo de documento móstranse no **Anexo A**.
- As páxinas de prensa poderán dixitalizarse en escala de grises (256 bits). No caso de que as cabeceiras seleccionadas para a dixitalización teñan portadas ilustradas ou entre as súas páxinas aparezan ilustracións e/ou fotografías, a obtención destas imaxes farase nunha escala de 16,7 millóns de cores (24 bits).

### 3.2. Identificación das imaxes

Todas as imaxes obtidas de proxectos de dixitalización sistemáticos terán que identificarse de forma unívoca para o que se recomenda cumprir coas seguintes pautas:

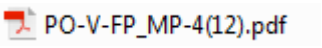
Para as publicacións non seriadas

a) O primeiro elemento na identificación das imaxes será o código asignado a institución depositaria dos orixinais
b) O seguinte elemento será a sinatura que identifica a obra
c) A continuación identificarase o número correlativo da imaxe na estrutura do obxecto dixital
d) Por último identificarase a extensión do formato de imaxe correspondente
e) Os primeiros 3 elementos irán separados entre si por un guión baixo, mentres que a extensión do ficheiro irá precedido por un punto

Exemplos

| | |
|---|---|
| ES-ScBG_PB4868_00008.tif | **ES-ScBG** → Biblioteca de Galicia<br>**PB4868** → O divino sainete : poema en oito cantos / M. Curros Enriquez<br>**00008** → número correlativo da imaxe<br>**.tif** → extensión do arquivo |

| | |
|---|---|
| LU-M-SM_e2-116_00728.jpg | **LU-M-SM** → Biblioteca do Seminario Santa Catalina. Diócese de Mondoñedo-Ferrol<br>**PB4868** → Adagiorum Des. Erasmi Roterodami Chiliades quatuor cum sesquicenturia cum diligentia ...<br>**00728** → número correlativo da imaxe<br>**.jpg** → extensión do arquivo |

| | |
|---|---|
|  | **PO-V-FP_MP-4(12)** → Biblioteca da Fundación Penzol<br>**MP-4(12)** → Carta de la Ría y Puerto de La Coruña número correlativo da imaxe [1]<br>**.pdf** → extensión do arquivo |

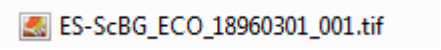
[1] Neste documento enténdese que os ficheiros PDF que se entreguen nun proxecto de dixitalización sistemática serán habitualmente multipáxina, polo que neste caso non será necesario identificar os números correlativos de cada imaxe. Podería existir un terceiro elemento neste tipo de arquivos si polo volume da obra se considerase pertinente a entrega do ficheiro PDF dividido en partes.

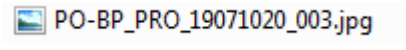
## Para as publicacións seriadas

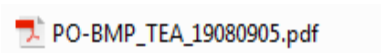
a) O primeiro elemento na identificación das imaxes será o código asignado a institución depositaria dos orixinais
b) O seguinte elemento será un código, de ata 5 caracteres, que identifique o título da publicación seriada
c) A continuación identificarase o ano, mes e día do número da publicación. No caso de existir erros de imprenta na data do ítem dixitalizado, a data seleccionada para a codificación das imaxes será sempre a da data real de publicación [1] [2]
d) O cuarto elemento será o número correlativo da imaxe na estrutura do obxecto dixital
e) Por último identificarase a extensión do formato de imaxe correspondente
f) Os primeiros 4 elementos irán separados entre si por un guión baixo, mentres que a extensión do ficheiro irá precedido por un punto

[1] Se non existise mes e/ou día concreto no ítem dixitalizado (ex: cabeceiras bimensuais ou semestrais) optarase polo día 01 e polo mes que máis se aproxime a súa data real de publicación
[2] Cando a data do ítem dixitalizado non sexa o suficientemente clara para diferenciar os exemplares da publicación entre si poderase engadir o número do ítem correspondente

## Exemplos

| | |
|---|---|
|  | **ES-ScBG** → Biblioteca de Galicia<br>**ECO** → El Eco de Santiago: diario independiente<br>**18960301** → Ítem do día 1 de marzo de 1896<br>**001** → número correlativo da imaxe<br>**.tif** → extensión do arquivo |

| | |
|---|---|
|  | **PO-BP** → Biblioteca Pública de Pontevedra<br>**PRO** → El Progreso: semanario independiente<br>**19071020** → Ítem do día 20 de outubro de 1907<br>**003** → número correlativo da imaxe<br>**.jpg** → extensión do arquivo |

| | |
|---|---|
|  | **PO-BMP** → Biblioteca-Museo de Ponteareas<br>**TEA** → El Tea: semanario independiente<br>**19080905** → Ítem completo do día 5 de setembro de 1908<br>**.pdf** → extensión do arquivo |

En todo caso, a Biblioteca de Galicia tratará de facilitar os códigos das institucións da memoria, as sinaturas de cada unha das obras dixitalizadas e os códigos que identifican aos títulos de publicacións seriadas.

### 3.3. Control de calidade

O control de calidade sobre os resultados do proceso de dixitalización é unha tarefa fundamental. Deberán revisarse todas as imaxes e prestarase atención, sobre todo, aos seguintes aspectos:

- Non existen imaxes borrosas
- Non hai imaxes repetidas
- As imaxes están perfectamente encadradas e verticais
- Os bordes das imaxes están correctamente recortados
- Están dixitalizadas a totalidade das páxinas
- A orde dos obxectos dixitais e das imaxes é o correcto
- Os ficheiros (en tódolos formatos) son perfectamente lexibles

### 3.4. Proceso de OCR (*Optical Character Recognition*)

A dixitalización de monografías e de publicacións periódicas que queiran formar parte do proxecto da Memoria Dixital de Galicia terá que levar asociado un proceso de recoñecemento óptico de caracteres (OCR) sobre toda as páxinas.

No caso das monografías, o resultado do OCR se proporcionara mediante un ficheiro PDF con OCR oculto por cada título dixitalizado. Os ficheiros PDF tamén estarán provistos dos marcadores necesarios para permitir o acceso polos diferentes niveis de contido (capítulos, seccións e epígrafes). Dentro da estrutura dos ficheiros METS de monografías incluiranse os elementos e as direccións necesarias dos títulos dixitalizados en formato PDF.

Para as publicacións seriadas o resultado do OCR proporcionarase en ficheiros normalizados XML segundo o esquema ALTO *(Analyzed Layout and Text Object)*, mantido pola Network Development and MARC Standards Office da Biblioteca do Congreso dos Estados Unidos. Os ficheiros XML deberán incluír unha estrutura xerárquica cos niveis de contido apropiados (tomos, números, páxinas, columnas, seccións, epígrafes e ilustracións) e incluirán por cada división de páxina un vínculo co ficheiro de imaxe correspondente.

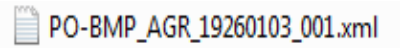
O resultado do OCR para este tipo de publicacións tamén se proporcionará mediante un ficheiro PDF con OCR oculto por cada número de prensa histórica dixitalizado. Os ficheiros PDF multipáxina estarán provistos, á súa vez, dos marcadores necesarios para permitir o acceso polos diferentes niveis de contido (números, seccións e epígrafes).

Na estrutura dos ficheiros METS de publicacións seriadas incluiranse os elementos e as direccións necesarias tanto para identificar os números en formato PDF como as páxinas co proceso de OCR codificado en ficheiros XML ALTO.

### 3.5. Identificación dos ficheiros ALTO-XML

Para a identificación de cada un dos arquivos ALTO-XML obtidos do proceso de cada una das páxinas das publicacións seriadas dixitalizadas recoméndase seguir o mesmo esquema que se detalla para as imaxes deste tipo de material no punto 3.2 deste documento. A diferencia virá determinada polo formato do arquivo que será **.xml**.

Exemplo

| | |
|---|---|
|   | **PO-BMP** → Biblioteca-Museo de Ponteareas<br>**AGR** → El Agro Celta<br>**18960301** → Ítem do día 3 de xaneiro de 1926<br>**001** → número correlativo do ficheiro ALTO-XML<br>**.xml** → extensión do arquivo |

# 4. CODIFICACIÓN E ESTRUTURA DOS DIRECTORIOS

## 4.1. Arquivos de preservación

A Subdirección Xeral de Bibliotecas, seguindo as normas e recomendacións internacionais, decidiu que os arquivos destinados á preservación nos proxectos de dixitalización sistemática vinculados á Memoria Dixital de Galicia, sexan obrigatoriamente os seguintes:

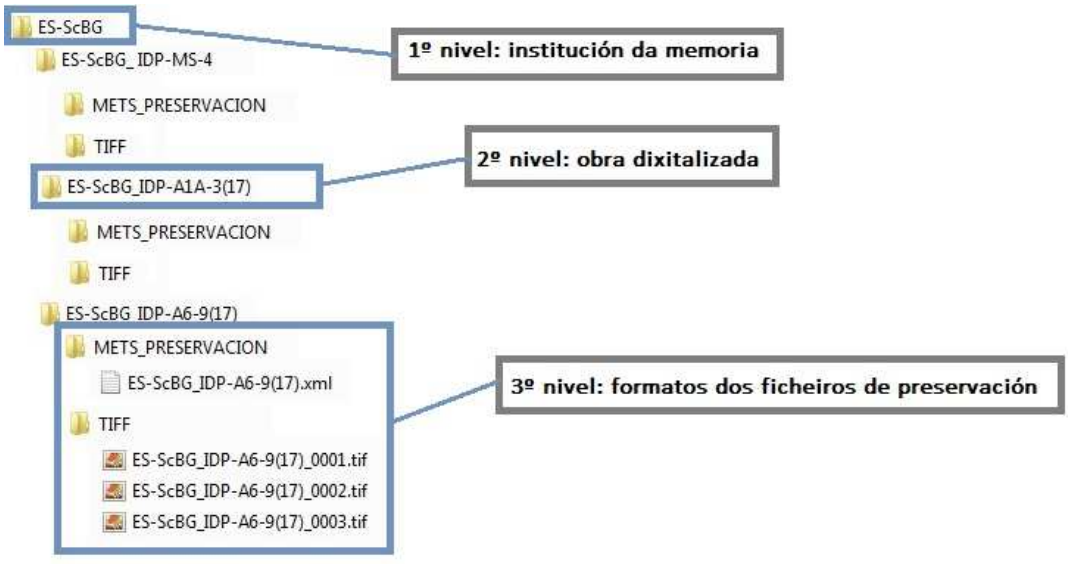
- Arquivos máster TIFF editados
- Ficheiros METS de preservación
- De xeito complementario tamén se poderá solicitar a preservación dos arquivos máster TIFF en cru para poder garantir o proceso completo de obtención das imaxes en relación co sinalado no punto 3.1. deste documento, isto é: *"no caso de que o fondo seleccionado permita a súa dixitalización a dobre páxina (manuscritos, monografías e incunables), do proceso de escaneado obterase un ficheiro TIFF máster e un TIFF recortado en dúas partes, é dicir, un ficheiro por cada páxina"*.

### Materiais non seriados

As imaxes obtidas como resultado do proceso de dixitalización seleccionadas para a preservación en formato TIFF, así como os ficheiros de metadatos asociados, almacenaranse seguindo os niveis de directorios establecidos a continuación:

1º nivel – Directorio xeral do proxecto por institución da memoria
2º nivel – Directorios de cada obra dixitalizada
3º nivel – Directorios por formatos dos ficheiros de preservación

Exemplo

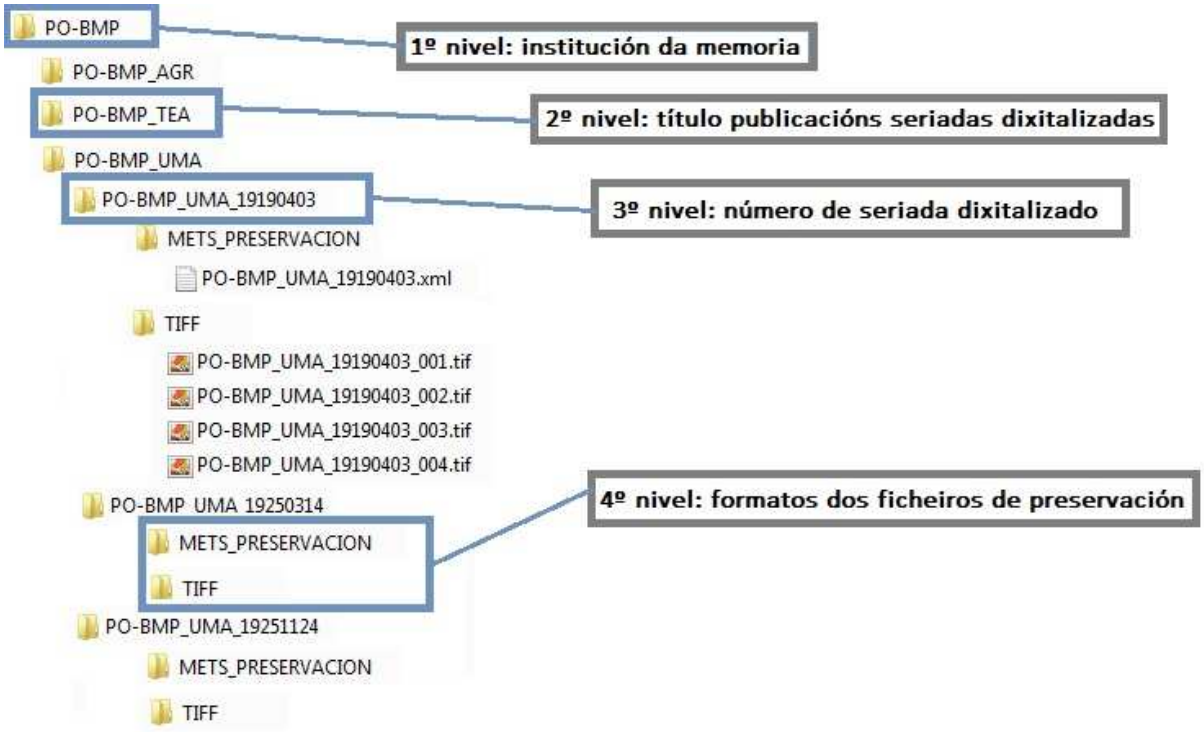


## Materiais seriados

As imaxes de publicacións seriadas seleccionadas para a preservación en formato TIFF, así como os ficheiros de metadatos asociados, almacenaranse seguindo os niveis de directorios establecidos a continuación:

1º nivel – Directorio xeral do proxecto por institución da memoria
2º nivel – Directorios por cada título de publicación seriada dixitalizada
3ª nivel – Directorios por cada número de seriada dixitalizado
4º nivel – Directorios por formatos dos ficheiros de preservación

Exemplo




## 4.2. Arquivos de difusión

Para a súa carga no repositorio de Galiciana-Biblioteca Dixital de Galicia, a Subdirección Xeral de Bibliotecas decidiu que os arquivos destinados á difusión nos proxectos de dixitalización sistemática vinculados á Memoria Dixital de Galicia serán os seguintes:

- Imaxes en formato JPG
- Ficheiros PDF con OCR oculto
- Miniaturas
- Ficheiros ALTO-XML (só nos proxectos de dixitalización de publicacións seriadas)

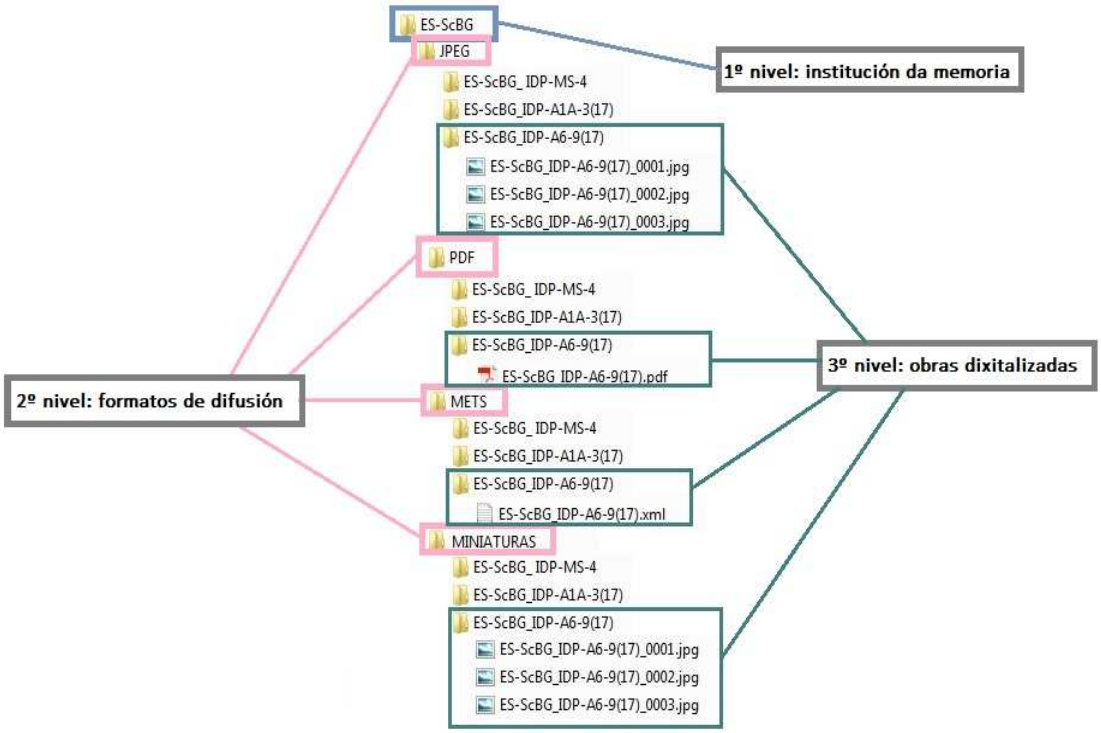
- Ficheiros METS de carga

## Materiais non seriados

Os arquivos destinados á difusión nos proxectos de dixitalización de materiais non seriados deberán entregarse en directorios coa seguinte codificación e estrutura:

1º nivel – Directorio xeral do proxecto por institución da memoria
2º nivel – Directorios por formatos
3º nivel – Directorios de cada obra dixitalizada

Exemplo




## Materiais seriados

Os ficheiros de difusión obtidos como resultado de procesos de dixitalización sistemática de prensa histórica almacenaranse seguindo os niveis de directorios establecidos a continuación:

1º nivel – Directorio xeral do proxecto por institución da memoria
2º nivel – Directorios por formatos
3º nivel – Directorios por cada título publicación seriada dixitalizada
4º nivel – Directorios por anos
5º nivel – Directorios por cada número de publicación seriada dixitalizado

## Exemplo


PO-BMP
ALTO
PO-BMP_AGR
PO-BMP_TEA
1908
PO-BMP_TEA_19080905
PO-BMP_TEA_19080905_001.xml
PO-BMP_TEA_19080905_002.xml
PO-BMP_TEA_19080912
PO-BMP_TEA_19080920
1909
1910
PO-BMP_UMA
JPEG
PO-BMP_AGR
PO-BMP_TEA
1908
PO-BMP_TEA_19080905
PO-BMP_TEA_19080905_001.jpg
PO-BMP_TEA_19080905_002.jpg
PO-BMP_TEA_19080912
PO-BMP_TEA_19080920
1909
1910
PO-BMP_UMA
MINIATURAS
PO-BMP_AGR
PO-BMP_TEA
1908
PO-BMP_TEA_19080905
PO-BMP_TEA_19080905_001.jpg
PO-BMP_TEA_19080905_002.jpg
PO-BMP_TEA_19080912
PO-BMP_TEA_19080920
1909
1910
PO-BMP_UMA
PDF
PO-BMP_AGR
PO-BMP_TEA
1908
PO-BMP_TEA_19080905
PO-BMP_TEA_19080905.pdf
PO-BMP_TEA_19080912
PO-BMP_TEA_19080920
1909
1910
PO-BMP_UMA
METS
PO-BMP_AGR
PO-BMP_TEA
1908
PO-BMP_TEA_19080905
PO-BMP_TEA_19080905.xml
PO-BMP_TEA_19080912
PO-BMP_TEA_19080920
1909
1910
PO-BMP_UMA
1º nivel: institución da memoria
2º nivel:formatos de difusión
3º nivel: título publicación seriada dixitalizada
4º nivel: anos dixitalizados
5º nivel: número de publicación seriada dixitalizado

## 5. METADATOS

Os metadatos son o conxunto de informacións que representan ao obxecto dixital. Son as ferramentas que nos permiten especificar a información contextual asociada a cada documento dixital, isto é, o seu contido, os formatos e características de cada ficheiro, o historial tanto da xeración como das transformacións sufridas por cada copia dixital e os diferentes hardwares e softwares vinculados coa súa creación, edición e preservación.

A información dos obxectos dixitais do proxecto da Memoria Dixital de Galicia ten que estar codificada seguindo o esquema METS. Cada unha das obras dixitalizadas dentro deste proxecto terá que estar asociada a dous arquivos METS, un deles destinado a DIFUSIÓN e outro a PRESERVACIÓN.

Os ficheiros METS de carga están configurados para a súa inxesta no repositorio de Galiciana-Biblioteca Dixital de Galicia, mentres que os ficheiros METS de preservación están destinados a súa importación no sistema de preservación da Memoria Dixital de Galicia. Ambos os dous arquivos deberán seguir as especificación que se expoñen a continuación.

### 5.1. Ficherios METS para a inxesta no repositorio de Galiciana-BDG

Os perfís técnicos que deberán seguirse para os ficheiros METS de carga en Galiciana-BDG son diferentes en función do tipo de material seleccionado para a súa dixitalización.

- **Ficheiros METS para a carga de monografías, manuscritos, material cartográfico e material gráfico**

As seguintes especificacións teñen como obxecto orientar nos metadatos asociados aos recursos orixinados a partir da dixitalización de monografías, manuscritos, material cartográfico e material gráfico (debuxos, fotografías, ...).

Os orixinais destes tipos de materiais que queiran formar parte do proxecto da Memoria Dixital de Galicia a través de Galiciana-BDG terán que estar descritos obrigatoriamente en formato MARC 21 bibliográfico. Ademais é recomendable que tanto a información da institución depositaria do orixinal dixitalizado como a da localización concreta do ítem estea codificada en formato MARC 21 de fondos e localizacións. De non ser así se pode incluír esa información no campo 852 do rexistro bibliográfico (subcampo $a para o código da institución e subcampo $j para a localización concreta do ítem).

A conversión deses rexistros bibliográficos e de fondos e localizacións ao esquema MARC-XML serán os **metadatos descritivos** do obxecto dixital que se embeberán no arquivo METS de carga.

Os obxectos dixitais derivados da dixitalización de calquera tipo de publicación non seriada

irán asociados a ficheiros METS de carga que terán como perfil de referencia o *Perfil METS da Biblioteca Virtual de Patrimonio Bibliográfico* accesible en :

- ✔ http://www.loc.gov/standards/mets/profiles/00000044.xml

Os requirimentos mínimos que deben cumprir os ficheiros METS de carga en Galiciana-BDG para este tipo de materiais son os seguintes:

- Incluír as direccións correctas aos namespaces e aos schemas empregados e deben estar ben formados e validados. A validación farase automaticamente a través dun software de validación de obxectos dixitais como JHOVE (JSTOR/Harvard Object Validation Environment) ou similar.
- Levar unha sección de metadatos descritivos <dmdSec> que embeba o rexistro bibliográfico en formato MARC-XML e, a continuación, o rexistro de exemplar en formato MARC-XML de fondos. Poderase substituír o rexistro MARC-XML de fondos por a información do exemplar no campo 852 embebido no rexistro bibliográfico.
- Incluír unha segunda sección de metadatos descritivos <dmdSec> na que se relacione a imaxe que vai servir como icona representativa da obra (normalmente unha portada ou ilustración).
- Establécese como recomendable que cada un dos formatos de DIFUSIÓN (JPEG, MINIATURAS e PDF) se reflicta nun <fileGrp> independente da Sección Arquivo (fileSec), aínda que para a súa carga en Galiciana-BDG só se require como obrigatorio a inclusión dun elemento de agrupación <fileGrp> para as imaxes JPEG destinadas a difusión a través de Internet.
- En cada un dos elementos <fileGrp> da Sección Arquivo terá que incluírse un atributo USE, ao que se lle asignará os seguintes valores dependendo do formato de imaxe que agrupe. Para os ficheiros de difusión en JPEG o atributo USE levará o valor “reference”; para as MINIATURAS, USE="thumbnail" e para o PDF, USE="ocr dirty".
- No caso de que sobre o obxecto dixital de material non seriado se houbese realizado un proceso de OCR codificado en ALTO XML, os ficheiros resultantes agruparíanse nun <fileGRP> independente do <fileSec>, asignándoselle ao seu atributo USE o valor “ocr”.
- Cada elemento de agrupación <fileGrp> terá un elemento <file> por cada imaxe dixitalizada que compoña a copia dixital nos diferentes formatos. No elemento <file> indicaranse os atributos correspondentes ao formato (MIMETYPE), o seu tamaño (SIZE) e a data de creación (CREATED). Ademais, no elemento <Flocat> de cada elemento <file> especificarase o atributo LOCTYPE="URL" e o seu valor debe indicar o camiño (estrutura de directorios) e o nome do ficheiro imaxe.
- Levar un Mapa Estrutural <structMap>, con atributo TYPE="PHYSICAL", no que se inclúa a información de paxinación, é dicir, o número de páxina asociado a cada imaxe.
- No caso de que se quixese cargar en Galiciana-BDG a versión en PDF con OCR oculto do obxecto dixital terá que incluírse un segundo <structMap>.

- Os elementos <structMap> incorporan un atributo LABEL cun valor que coincide co título do material non seriado ao que fai referencia o obxecto dixital.

No **Anexo B** móstrase un exemplo comentado de arquivo METS de carga para monografías, manuscritos, material cartográfico e material gráfico.

- **Ficheiros METS para a carga de publicacións seriadas**

Os orixinais de números de publicacións seriadas seleccionados para formar parte do proxecto da Memoria Dixital de Galicia terán que estar descritos obrigatoriamente en formato MARC 21 bibliográfico.

A información de fondos e localizacións de cada un dos títulos de prensa irá nun rexistro independente seguindo o formato MARC 21 de fondos e localización. De non ser así, esa información poderá ir embebida no campo 852 do rexistro bibliográfico (subcampo $a para o código da institución e subcampo $j para a localización concreta do ítem).

A información de periodicidade e numeración de cada ítem de publicación seriada codificase seguindo o establecido para os campos **853-878 – Holdings Data – General Information** do MARC 21 de fondos e localizacións publicado pola Libray of Congress.

A conversión deses rexistros bibliográficos e de fondos e localizacións ao esquema MODS (Metadata Object Description Schema) serán os metadatos descritivos do obxecto dixital que se embeberán no arquivo METS de carga.

Os obxectos dixitais derivados da dixitalización de calquera tipo de publicación seriada teñen que ir asociados a ficheiros METS de carga que terán como perfil de referencia o *METS Profile for Historial Newspapers* accesible en :

- ✔ http://www.loc.gov/standards/mets/test/ndnp/00000010.xml

En particular, os METS para a carga de publicacións seriadas terán que seguir as seguintes especificacións:

- Incluirán as direccións correctas aos namespaces e aos esquemas empregados e deben estar ben formados e validados. A validación farase automaticamente a través dun software de validación de obxectos dixitais como JHOVE (JSTOR/Harvard Object Validation Environment) ou similar.
- Levarán unha sección de metadatos descritivos (dmdSec) que embeba o rexistro bibliográfico en formato MODS (Metadata Object Description Schema), con información normalizada de numeración e cronoloxía segundo MARC 21 de fondos e localizacións.
- Incluirán unha segunda sección de metadatos descritivos <dmdSec> na que se

relacione a imaxe que vai servir como icona representativa do ítem de prensa (normalmente a primeira páxina ou unha ilustración).

- Establécese como recomendable que cada un dos formatos de DIFUSIÓN (JPEG, MINIATURAS e PDF) se reflicta nun <fileGrp> independente da Sección Arquivo (fileSec), aínda que para a súa carga en Galiciana-BDG só se require como obrigatorio a inclusión dun elemento de agrupación <fileGrp> para as imaxes JPEG destinadas a difusión a través de Internet.
- En cada un dos elementos <fileGrp> da Sección Arquivo terá que incluírse un atributo USE, ao que se lle asignará os seguintes valores dependendo do formato de imaxe que agrupe. Para os ficheiros de difusión en JPEG o atributo USE levará o valor "reference"; para as MINIATURAS, USE="thumbnail" e para o PDF, USE="ocr dirty".
- No caso de obxectos dixitais resultado da dixitalización de números de publicacións periódicas é obrigatorio realizar un proceso de OCR codificado en ALTO XML. Os ficheiros resultantes agruparanse nun <fileGRP> independente do <fileSec>, asignándoselle ao seu atributo USE o valor "ocr".
- Cada elemento de agrupación <fileGrp> terá un elemento <file> por cada imaxe dixitalizada que compoña a copia dixital nos diferentes formatos. No elemento <file> indicaranse os atributos correspondentes ao formato (MIMETYPE), o seu tamaño (SIZE) e a data de creación (CREATED). Ademais, no elemento <Flocat> de cada elemento <file> especificarase o atributo LOCTYPE="URL" e o seu valor debe indicar o camiño (estrutura de directorios) e o nome do ficheiro imaxe.
- Terá un Mapa Estrutural (structMap), con atributo TYPE="LOGICAL", no que se inclúa a información de paxinación e que estableza a relación entre as imaxes JPG e os seu correspondente ficheiro ALTO XML.
- No caso de que se quixese cargar en Galiciana-BDG a versión en PDF con OCR oculto do obxecto dixital terá que incluírse un segundo <structMap>.
- Os elementos <structMap> incorporan un atributo LABEL cun valor que coincide co título do material non seriado ao que fai referencia o obxecto dixital.

No **Anexo C** móstrase un exemplo comentado de arquivo METS de carga para publicacións periódicas.

### 5.2. Ficheiros METS para a inxesta no repositorio de preservación da MDG

Todas as obras dixitalizadas que formen parte da Memoria Dixital de Galicia terán que levar asociado un arquivo METS de preservación para a súa inxesta no sistema de preservación dixital da Xunta de Galicia.

Aínda que o tema da preservación dixital a longo prazo é algo moi presente nos foros de bibliotecas, arquivos e museos desde hai anos, no noso país son escasos os proxectos levados a cabo para a preservación dos obxectos non nacidos en dixital. Unha das institucións de referencia neste tema é a Biblioteca Nacional de España que deseñou un primeiro perfil de metadatos METS-PREMIS para materiais dixitalizados dentro do seu

proxecto de dixitalización sistemática de materiais impresos en formato papel destinados a Biblioteca Digital Hispánica.

En termos xerais, as especificacións que se detallan a continuación para os ficheiros METS de preservación da Memoria Dixital de Galicia seguen o perfil establecido pola BNE. O obxectivo é recoller nun único ficheiro METS de preservación todos os metadatos descritivos, administrativos, técnicos e de preservación referentes aos arquivos que se xeneran despois da dixitalización dunha obra. Tal e como se expón no punto 4 deste documento, para os procesos de dixitalización sistemática a Subdirección Xeral de Bibliotecas tomou a decisión de preservar:

- Arquivos máster TIFF editados
- De xeito complementario tamén se poderá solicitar a preservación dos arquivos máster TIFF en cru para poder garantir o proceso completo de obtención das imaxes en relación co sinalado no punto 3.1. deste documento, isto é: *"no caso de que o fondo seleccionado permita a súa dixitalización a dobre páxina (manuscritos, monografías e incunables), do proceso de escaneado obterase un ficheiro TIFF máster e un TIFF recortado en dúas partes, é dicir, un ficheiro por cada páxina"*.

A Subdirección Xeral de Bibliotecas considera que, polo momento, os formatos de difusión xa contan co respaldo suficiente para a súa preservación a longo prazo mediante as ferramentas e sistemas habilitados para o almacenamento dos obxectos dixitais no repositorio de Galiciana-BDG.

Hai que sinalar que na actualidade a configuración na estrutura dos directorios para a inxesta no repositorio de Galiciana-BDG e no repositorio de preservación da Memoria Dixital de Galicia non é coincidente. Por este motivo, a Subdirección Xeral de Bibliotecas considera que a entrega dos proxectos sistemáticos de dixitalización con dúas estruturas diferentes (unha para difusión e outra para preservación) en todos os formatos obtidos da copia dixital pode incrementar o custe do proxecto dun xeito non asumible polas institucións da memoria promotoras do mesmo. Para tratar de minimizar ese impacto económico, a Subdirección Xeral de Bibliotecas propón entregar unicamente nunha estrutura diferente os arquivos TIFF máster editados.

Os arquivos METS de preservación para a inxesta no repositorio de preservación da Memoria Dixital de Galicia deberán cumprir, como mínimo, as seguintes características:

- Incluirán os metadatos seleccionados pola Subdirección Xeral de Bibliotecas que describan os termos que se precisan para a preservación da copia dixital. Estes metadatos obtivéronse do Dicionario de Datos PREMIS, versión 3.0, denominado PREMIS Data Dictionary, e elaborado segundo o Modelo de Referencia do Open Archival Information System (OAIS), ISO14721.
- Das cinco entidades ás que pode referirse un arquivo de preservación segundo

PREMIS, nos ficheiros de preservación utilizaranse as seguintes: axente, obxecto e evento. Estas unidades semánticas estarán embebidas na sección de metadatos administrativos do ficheiro METS.

- Farán referencia a información necesaria para a preservación dos arquivos máster TIFF produto da dixitalización. A información deste formato de imaxe irá nun <fileGrp> da Sección Arquivo (fileSec) cos enlaces as imaxes correspondentes. O valor do atributo USE para os ficheiros máster será “arquive”.
- Incluirán unha sección de metadatos descritivos codificados, como mínimo, nos esquemas MARC 21 XML ou MODS, segundo o tipo de material, para a identificación dos obxectos dixitais.
- Entregaranse no ficheiro XML os metadatos de dereitos de autor e propiedade intelectual dos obxectos dixitais, estruturados segundo o esquema METSRigths. Neles detallaranse as características de dereito de uso das publicacións segundo determine a Memoria Dixital de Galicia.

No **Anexo D** móstrase un exemplo comentado de arquivo METS para a inxesta no repositoiro de preservación da Memoria Dixital de Galicia.

BIBLIOGRAFÍA

*Directrices para proyectos de digitalización de colecciones y fondos de dominio publico, en particular para aquellos custodiados en bibliotecas y archivo*s
http://hdl.handle.net/10421/3342

*Requisitos técnicos de los proyectos de digitalización de* patrimonio *bibliográfico y de prensa histórica de la SGCB*
http://hdl.handle.net/10421/8981

*Proceso de digitalización en la Biblioteca Nacional de España. Biblioteca Digital Hispánica*
http://www.bne.es/webdocs/Catalogos/ProcesoDigitalizacionBNE.pdf