1. CONTEXTO

El Museo del Prado abrió al público por primera vez en el año 1819 (aunque el edificio fue diseñado en 1785), mostrando un catálogo inicial de 311 pinturas. Las colecciones más emblemáticas del Museo se gestaron durante el S. XVI, bajo el mandato de Carlos V, y se fueron creando más obras durante el reinado de monarcas posteriores (tanto Austrias como Borbones), de autores tan célebres como El Greco, Goya, Tiziano o Velázquez.

Durante los Siglos XIX y XX el número de visitantes del Museo del Prado se incrementó de forma notable, hasta albergar en la actualidad más de **19.000 obras** de distintos autores a lo largo de diversas épocas históricas. Debido a la gran cantidad de obras y visitantes que recibe el Museo, el edificio ha tenido que acometer diversas ampliaciones para dar cabida a la demanda de visitantes.

Una de las características del museo es su página web (https://www.museodelprado.es/), donde se pueden consultar todas sus obras, clasificadas por autor, temática, época, orden alfabético... Uno de los aspectos más llamativos es que se puede consultar la **ficha técnica** de cada obra, donde se detallan las características principales del cuadro, lo cual ha facilitado realizar el web scrapping de la página web del Museo del Prado.

2. TÍTULO

Ficha Técnica (Características) de las obras del Museo del Prado

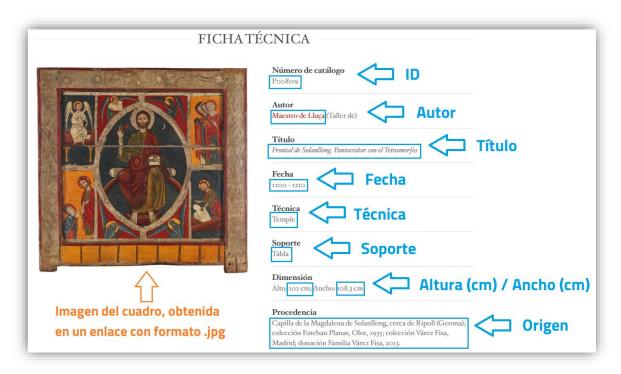
3. DESCRIPCIÓN DEL DATASET

En la web del Museo del Prado existen más de 19.000 obras de arte que se pueden consultar todas en el enlace https://www.museodelprado.es/coleccion/obras-de-arte. Inicialmente, la imagen estática de la web muestra 18 obras, y a medida que se va haciendo scroll down, va cargando otras 18 obras adicionales, así hasta completar las 19.605 obras totales que alberga la web.

La intención era extraer la información de la totalidad de las obras, sin embargo, debido a las limitaciones de scrapping, "sólo" se han podido extraer poco menos de la mitad (9.680 obras). Una vez obtenidos los 9.680 enlaces, se han almacenado en una lista y luego se han analizado de forma individual para poder extraer los datos de la ficha técnica de cada uno.

4. REPRESENTACIÓN GRÁFICA

En la URL individual de cada obra, se puede encontrar la ficha técnica del cuadro, que consta de la siguiente información y es la que se ha utilizado para realizar el web scrapping:



5. CONTENIDO.

El dataset extraído de la web consta de:

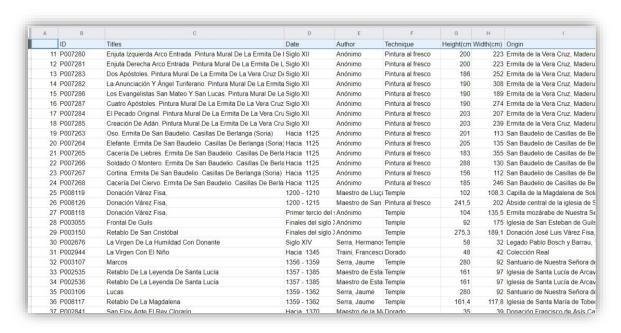
- 9.680 observaciones (filas)
- 9 atributos (columnas).
 - ID: número de catálogo que otorga el Museo a cada obra. Este número es único e irrepetible, para poder identificar fácilmente a cada obra.
 - o **Título**: el título de la obra.
 - Fecha: año de creación de la obra. Debido a la antigüedad de las obras, algunas de ellas tienen como fecha "Siglo XII" o "Antes de 1489" (por ejemplo), ya que es muy difícil definir un año concreto para las mismas.
 - o Autor: persona legítima encargada de realizar la obra.
 - Técnica: tipo de pintura o técnica para realizar la obra (pintura al fresco, óleo, temple, papel a la sal, albúmina...)
 - o Altura: altura del cuadro en centímetros (cm).
 - o **Ancho**: ancho del cuadro en centímetros (cm).
 - Soporte: el soporte pictórico es la base sobre la que se desarrolla la obra, es decir, la superficie sobre la que se aplica el color en una pintura. Pueden ser: tabla, lienzo, mural...

 Procedencia: lugar de origen donde se concibió la obra. Dependiendo de la tipología de la obra, se puede mencionar solo la ciudad o a veces se puede concretar incluso la ermita, iglesia o edificio donde se creó.

Había un décimo atributo, llamado "**Serie**", que especificaba a qué serie o conjunto de obras pertenecía cada obra. Por ejemplo, un cuadro podría pertenecer a una serie de 15 cuadros de un mismo autor u origen. Pero, debido a que este atributo lo tenían muy pocas obras (la mayoría eran *missing values* (NaN)), se decidió eliminar esta característica.

También cabe mencionar que hay una primera característica que se ha utilizado con el único propósito de numerar (#) de forma interna las obras que se han ido analizando, desde la 0 hasta la 9.680.





6. AGRADECIMIENTOS

Antes de empezar a trabajar, a través de la página de contacto https://www.museodelprado.es/contacto, se escribió un correo a la dirección oficial del Museo del Prado web@museodelprado.es/contacto, informando que se iba a realizar web scrapping en su página web con fines únicamente académicos y didácticos, y solicitamos permiso para ello. Sin embargo, nunca llegamos a obtener respuesta por parte del web@museodelprado.es/contacto, informando que se iba a realizar web scrapping en su página web con fines únicamente académicos y didácticos, y solicitamos permiso para ello. Sin embargo, nunca llegamos a obtener respuesta por parte del web@museodelprado.es/contacto, informando que se iba a realizar web scrapping en su página web con fines únicamente académicos y didácticos, y solicitamos permiso para ello. Sin embargo, nunca llegamos a obtener respuesta por parte del web.//www.museodelprado.es/, informando que se iba a realizar web scrapping en su página web con fines únicamente académicos y didácticos, y solicitamos permiso para ello. Sin embargo, nunca llegamos a obtener respuesta por parte del web.//www.museodelprado.es/, informando que se iba a realizar web

Contacto

Centro de Atención al Visitante cav@museodelprado.es

Contacto general museo.nacional@museodelprado.es

Biblioteca biblioteca@museodelprado.es

Comunicación (Prensa y video) area.comunicacion@museodelprado.es

Redes sociales redes@museodelprado.es

Web web@museodelprado.es

Colección Prado consultas.coleccion@museodelprado.es

A pesar de no obtener respuesta, se procedió a trabajar en el proyecto, ya que toda la información extraída es pública, y en ningún momento tuvimos que registrarnos u obtener una clave de usuario específica, de modo que no era posible incurrir en posibles problemas legales. Por ello, agradecemos a la institución del Museo del Prado poner a nuestra disposición toda su información pública; damos por hecho que, si esto les supusiera algún problema, se hubieran puesto en contacto con nosotros. Asumimos la ausencia de respuesta como una autorización a nuestro proyecto.

De todas las obras que alberga el Museo, solo se extrajeron la mitad de enlaces, para no sobrecargar de peticiones la web del mismo. Una vez extraída toda la información, sólo se utilizó la misma con fines didácticos y académicos, como ya se ha mencionado, y nunca se han utilizado ni se utilizarán con fines comerciales o con ánimo de lucro.

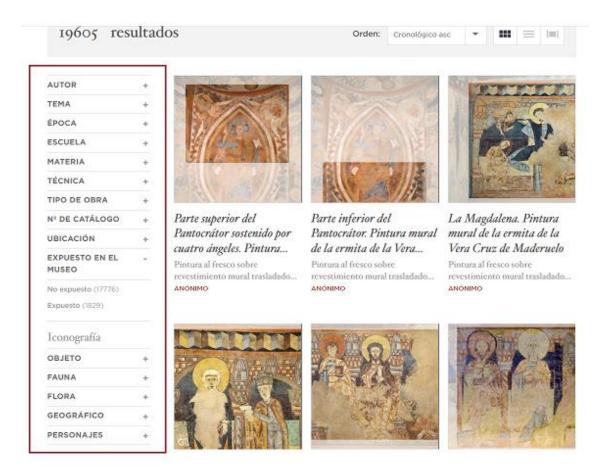
7. INSPIRACIÓN

Debido a la gran cantidad de obras del Museo, consideramos interesante el dataset que hemos generado, ya que con una simple consulta se puede encontrar la información relativa a cualquier cuadro.

Aunque la web facilita la clasificación de sus obras, y éstas se pueden consultar por Autor, Tema, Época, Técnica, etc. y también por iconografía (cuadros que

representen flora, fauna, armas, alimentos, etc), la página principal donde se exponen las 19.605 obras de arte resulta un tanto impracticable, ya que las obras se cargan mediante scroll down, y a medida que van apareciendo más cuadros, más difícil resulta navegar por la página (y en consecuencia se satura el navegador).

Así pues, el dataset resulta útil para albergar la información de TODOS los cuadros en un solo documento, y no tener que fragmentarlo en múltiples secciones o enlaces diferentes.



Así pues, en el dataset se puede consultar cualquier aspecto relativo a la ficha técnica de cada cuadro, tales como el título, el autor, la fecha o la técnica de la obra. También se incluyen las URLs de las imágenes de cada obra, en caso que se quiera visualizar la imagen a la que hace referencia.

8. LICENCIA

Hemos utilizado la licencia Creative Commons Attribution (CC4.0), ya que abarca tanto material original del autor como material generado a partir de otras fuentes y modificado por el autor.

9. CÓDIGO

El código implementado en este trabajo esta disponible en el repositorio de GitHub:

https://github.com/llcassales/Scrapping

10. DATASET

El dataset obtenido luego de aplicar web-scrapping en la web del Museo del Prado esta disponible en Zenodo https://doi.org/10.5281/zenodo.6433134.

11. VIDEO

El vídeo lo hemos grabado por separado, y luego se han juntado las dos partes mediante un programa de edición de vídeo (Adobe Premiere Pro). El archivo .mp4 resultante se ha colgado en Google Drive (se han otorgado permisos de visualización a los dos integrantes y a la profesora Mireia Calvo González), este es el enlace:

https://drive.google.com/file/d/1wzYHZpAhzOK2TSrMvDeB0EvPUwoPyTkA/view

Contribuciones			Firma
Investigación previa			LS, AR
Redacción	de	las	LS, AR
respuestas			
Desarrollo del código			LS, AR