

PATHFINDER

Herramienta para el análisis
gráfico y textual de datos
procedentes de ExpoFinder

Guía de uso preliminar abreviada
v1.0.0 2018 ES

Universidad de Málaga. Departamento de Historia del Arte. Grupo
de investigación iArHis_LAB

Antonio Cruces Rodríguez

AUTOR: Antonio Cruces Rodríguez.

ORCID ID: orcid.org/0000-0001-6376-9608.

EDITA: Grupo de investigación iArtHis_LAB. Universidad de Málaga.

Revisión: 1

Fecha: 20180831

Revisor: Antonio Cruces Rodríguez, antonio.cruces@uma.es



Esta obra está sujeta a una licencia Creative Commons:
Reconocimiento - No comercial - SinObraDerivada (cc-by-nc-nd):
<http://creativecommons.org/licenses/by-nc-nd/3.0/es>. Cualquier
parte de esta obra se puede reproducir sin autorización, pero
con el reconocimiento y atribución de los autores. No se puede
hacer uso comercial de la obra y no se puede alterar,
transformar o hacer obras derivadas.

TABLA DE CONTENIDO

PRESENTACIÓN	6
INTERFAZ DE USUARIO	7
FUNCIONAMIENTO	9
Ajustes	9
DATOS	13
Panel	13
Filtro y búsqueda	14
Listado	17
Criba	18
Relaciones	21
Gráficos	21
Redes	24
Mapas	25
OTRAS CONSIDERACIONES	25

TABLA DE TÉRMINOS ESPECÍFICOS, ABREVIATURAS Y ACRÓNIMOS

BD	Base de datos
CSV	Formato de exportación. Archivo de texto simple con los campos separados por el carácter “coma” (,). Compatible con la mayoría de hojas de cálculo
EF	ExpoFinder
GiB	Gibibyte. Unidad de medida de información según el estándar ISO/IEC 80000-13. 1 GiB = 1.073.741.824 B = 2^{30} bytes
KISS	Acrónimo de <i>Keep It Simple, Stupid</i> , principio atribuido a Kelly Johnson, ingeniero jefe de la división de investigación de Lockheed
MiB	Mebibyte. Unidad de medida de información según el estándar ISO/IEC 80000-13. 1 MiB = 1.048.576 B = 2^{20} bytes
OSM	Acrónimo de <i>OpenStreetMap</i> . Compañía sin ánimo de lucro dedicada a la distribución de sistemas de geoposicionado y trazado de mapas
PF	Pathfinder
RAM	Acrónimo de <i>Random Access Memory</i> . Memoria volátil
stack trace	Trazado de la pila. Término de uso técnico utilizado en depuración de errores de aplicación
<i>sticky footer</i>	“pie adhesivo”; pie de página que permanece fijo, independientemente de la posición de la barra de desplazamiento (<i>scroll</i>)
TAB	Formato de exportación. Archivo de texto simple con los campos separados por el carácter de tabulación. Compatible con la mayoría de hojas de cálculo
XML	Acrónimo de <i>Extended Markup Language</i> . Formato de exportación.

TABLA DE ILUSTRACIONES

Ilustración 1. Zonas de pantalla.	7
Ilustración 2. “limpieza” de “Local Storage”.	13
Ilustración 3. Tabla de combinaciones en el proceso de filtrado.	16
Ilustración 4. Ejemplo de condiciones de filtrado.	16
Ilustración 5. Gráficos.	22

PRESENTACIÓN

PATHFINDER (PF) es un desarrollo software realizado por el grupo de investigación iArtHis_LAB, del Departamento de Historia del Arte de la Universidad de Málaga (España), como parte de los proyectos Exhibitium FBBVA 2014 y ArtCatalogs (HAR2014-51915-P) para poner a disposición de los investigadores interesados en la aplicación del paradigma digital al estudio de las Humanidades (en especial de la Historia del Arte) una herramienta versátil que permita realizar análisis cuantitativos de la información cualitativa propia de la disciplina.

El objetivo de PF es mostrar las posibilidades que aporta el tratamiento geográfico de los datos sobre exposiciones temporales de contenido artístico y sobre los catálogos recopilados ambos tipos de registro en ExpoFinder, utilizando diversas técnicas para representar las densidades de entidades organizadoras, financiadoras o promotoras, de centros expositivos, de actores (artistas, comisarios, ilustradores, editores, ...) y las relaciones entre todos esos elementos.

La aplicación está concebida para poder localizar ubicaciones y realizar estudios comparativos entre ellas. Es posible filtrar los datos por años de comienzo de las exposiciones, por número de eventos organizados, por localizaciones y por otras características.

El equipo de desarrollo estará encantado de atender cualquier consulta que desee dirigirle, preferentemente utilizando como medio el correo electrónico. Por favor, en el mensaje que nos envíe asegúrese de incluir como referencia su nombre de usuario ExpoFinder y, si es posible, la fecha de la descarga de datos más reciente.

Nuria Rodríguez Ortega, Antonio Cruces Rodríguez
Universidad de Málaga. Facultad de Filosofía y Letras.
Departamento de Historia del Arte.
Campus de Teatinos. Bulevar Louis Pasteur, s/n.
29071 Málaga (España).
Teléfono: +34 952 131 690.
Correo electrónico: nro@uma.es, antonio.cruces@uma.es.

INTERFAZ DE USUARIO

La pantalla se divide en zonas, con la particularidad de que las denominadas F y G ahora forman un bloque fijo, siempre “pegado” al pie de la pantalla (*sticky footer*), de manera que la zona E se desplaza “por debajo” de dicho pie.

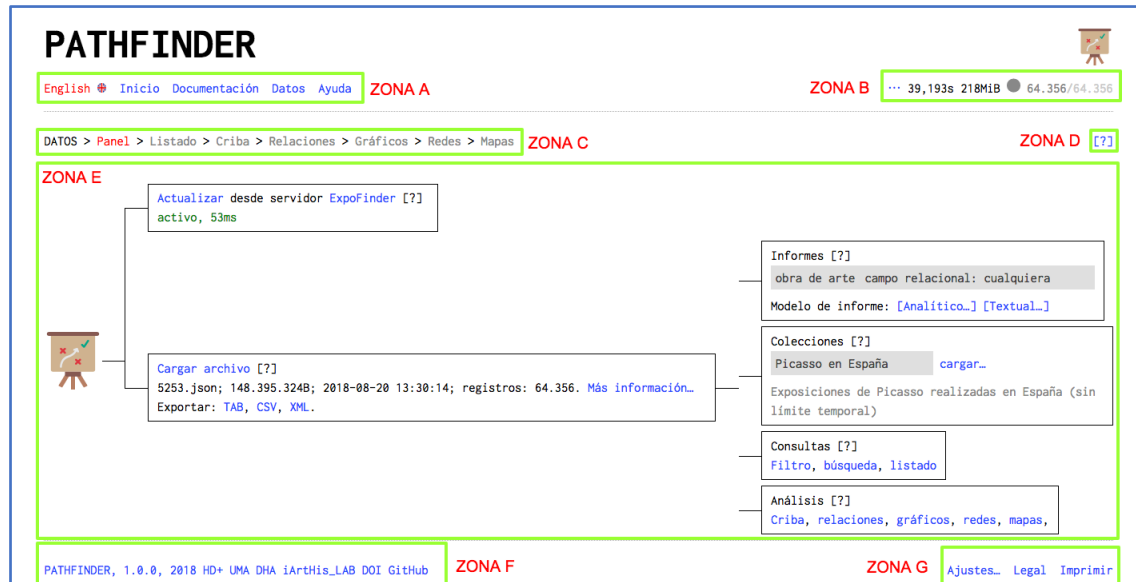


Ilustración 1. Zonas de pantalla.

Las zonas son las siguientes:

1. MENÚ PRINCIPAL. Contiene las siguientes opciones:
 - a. Link de cambio de idioma. Muestra el idioma al que se va a cambiar (no el que está activo) y su icono correspondiente.
 - b. Inicio (página principal). Desde ella se puede acceder a Primeros pasos (en desarrollo). En este apartado se ha añadido una prestación en el submenú ¿Cómo funciona?, dedicado al interfaz de usuario: cuando se alude a una parte de la pantalla, al pulsar sobre el enlace que la menciona se muestra dicha zona desplazando el conjunto y marcándola de forma intermitente.
 - c. Documentación. En desarrollo.
 - d. Datos. Pantalla que se describe más adelante.
 - e. Ayuda. Despliega ventana emergente con opciones de información y ayuda general de la aplicación; la ayuda específica se muestra con cada menú y/o submenú, pero la general permanece invariable. Tiene las siguientes opciones:
 - i. Árbol de menús. Despliega de manera jerárquica las diferentes alternativas y operaciones de menús y submenús (en desarrollo; hay que adaptarlo a los cambios realizados).

- ii. Preguntas frecuentes y sus correspondientes respuestas. Irá ampliándose conforme sea necesario.
 - iii. Operatoria. Diagrama de flujo con los principios básicos de funcionamiento. En proceso de modificación.
 - iv. ¿Cómo...? Si deseo realizar X los pasos son...
2. INDICADOR DE ESTADO. Muestra la situación de PF en cada momento. Dependiendo de las circunstancias (archivo de datos cargado o no) puede exhibir unas informaciones u otras. De izquierda a derecha comprende los siguientes elementos.
- a. Marcador "...". Abre una ventana emergente que muestra el estado de carga de las librerías (bibliotecas de código) pertinentes en cada caso, y depende de qué opción de menú se haya seleccionado mostrará unas u otras cargadas o no. El grueso de librerías cargadas corresponde a la opción d (Datos).
 - b. Cronómetro. Muestra la duración en segundos de la operación de base de datos (BD) más reciente. Antes de cargar BD no muestra nada. Cuando se carga el archivo de datos muestra el tiempo total de preparación. En un ordenador normal (+4GiB RAM, Intel Core i3), y según los ajustes seleccionados (ver más adelante) oscilará entre 20s y 45s para un archivo de datos de alrededor de 150 MiB (unos 65.000 registros principales).
 - c. Memoria principal usada por BD. En MiB. Según los ajustes seleccionados (ver más adelante) oscilará entre 83MiB y 218MiB para un archivo de datos de alrededor de 150 MiB (unos 65.000 registros principales).
 - d. Información sobre el filtro. Muestra un micrográfico de tarta con el total de registros de BD y el porcentaje filtrado. Es un enlace, y al pulsar sobre él se muestra una ventana emergente con información sobre el filtro (registros filtrados por tipo) más enlaces directos a los procesos de filtro y búsqueda.
3. SUBMENÚ. Cambia según la opción de menú seleccionada. Ofrece las diferentes opciones encabezadas a la izquierda por la opción de menú activa y separadas por el símbolo ">".
4. AYUDA ESPECÍFICA. Aludida en la zona A, cambia según la operación activa en cada momento. En desarrollo.
5. ÁREA DE OPERACIONES. Es la zona activa de la pantalla, en la que se llevan a cabo todas las tareas principales de PF. Cuando se imprime la pantalla es la única que se muestra (las otras no aparecerán en el texto impreso). Se desplazará "por debajo" del pie fijo cuando sea necesario.

6. ESTADO DE LA APLICACIÓN Y RESPALDO ACADÉMICO. El nombre de PF y su versión son un enlace que muestra una ventana emergente con datos de copyright más información sobre el estado del navegador y los enlaces con los servidores (de datos y de aplicación). Los demás son enlaces a las respectivas instituciones.
7. AJUSTES, INFORMACIÓN LEGAL E IMPRESIÓN. Concentra esas tres operaciones.
 - a. Ajustes. La pantalla se subdivide en dos columnas.
 - i. A la izquierda se muestra información genérica de requisitos, cumplimiento de los mismos por parte del navegador (muchos de ellos no pueden ser detectados, como la memoria RAM o el microprocesador por razones de seguridad), estado de la memoria y velocidad de procesamiento calculada para cada dispositivo.
 - ii. A la derecha se muestran los ajustes. Al pulsar sobre un valor (son enlaces) se selecciona. Ver más adelante el significado de cada uno.
 - b. Legal. Muestra una pantalla con tres opciones:
 - c. Contacto. Indica la dirección para contactar con el soporte PF, así como las direcciones IP y física detectadas del dispositivo de acceso y el estado de carga de BD.
 - d. Normativa. Información legal sobre sujeción a normativas, cookies y derechos de copia.
 - e. Licencias. Tanto la propia de PF como las de los elementos utilizados en bibliotecas de código.
 - f. Imprimir. Como se ha mencionado solamente se imprime la zona E.

FUNCIONAMIENTO

La arquitectura de PF trata de atenerse a la simplicidad, al minimalismo gráfico y visual, intentando no emplear demasiados recursos en el interfaz que puedan distraer al usuario de la finalidad de la aplicación, y al principio KISS (*Keep It Simple, Stupid*), atribuido a Kelly Johnson, ingeniero jefe de la división de investigación de Lockheed, allá por la década de los 60 del siglo pasado. PF es una herramienta, no un “divertimento”. Hay que aprender a usarla, pero la sencillez de su funcionalidad facilita una curva de aprendizaje gradual y asequible.

Ajustes

Permiten controlar el comportamiento de las diferentes partes de la aplicación. Dispone de un enlace en la zona G de pantalla. Sus elementos principales son los siguientes:

1. Captura de errores. Si PF muestra o no una ventana emergente indicando que se ha producido un error.
2. Mensajes de error detallados. Indicar el origen del error (*stack trace*, trazado de la pila de instrucciones: cual fue la sucesión de instrucciones que dispararon el error). No es útil para el usuario sin conocimientos técnicos.
3. Mostrar error en consola de desarrollo. La mayoría de los navegadores disponen de una “consola” donde mostrar información e interactuar. Indica si los errores aparecerán o no en ella. No es útil para el usuario sin conocimientos técnicos.
4. Aviso antes de recargar la página. Si PF muestra un mensaje ANTES de recargar la página (solamente si se ha modificado algo en ella); al recargarla se perderían las operaciones realizadas con los datos en caso de que
5. Precarga de relaciones. Una de las más complejas tareas que realiza PF es el cálculo de vínculos entre los registros. Hay que tener en cuenta que un registro viene definido por su cuerpo de datos principal (ID, nombre, tipo de registro), sus metadatos (geográficos, temporales, ...) y sus taxonomías (clasificaciones). El registro de Picasso (ID NNNN) no es un único elemento, sino que de él cuelgan miles de metadatos y decenas de taxonomías. Si cada vez que hay que calcular relaciones es preciso llevar a cabo la consulta la velocidad de las prestaciones se degrada y PF responde más lentamente. Así que cabe la posibilidad de guardar en memoria dicho cálculo, realizándolo solo una vez durante la carga de BD; así, la carga es más lenta pero no el resto de las operaciones. Inconveniente: ocupa más memoria RAM (la diferencia entre 83MiB y 218MiB para 150MiB de archivo JSON y unos 65.000 registros). Por eso solo es aconsejable activar esta opción si se dispone de más de 8GiB de RAM.
6. Mapas estáticos generados por Google u OpenStreetMap. Un “mapa estático” es una imagen; no es “navegable”, y viene bien (ocupa menos RAM) para mostrar en listado la ubicación de un registro. Google, a pesar de las apariencias, es un proveedor de pago para este tipo de servicios, y limita el número de consultas que se pueden realizar a sus servidores; no así OpenStreetMap (OSM). La respuesta de Google es mucho más rápida, pero a veces (dependiendo de si cree que has superado el número de consultas admisible por unidad de tiempo) falla.
7. Formato z-score numérico (de 1 a 5) o gráfico (igual, pero usando ★ en vez de guarismos). La unidad tipificada (z-score, en argot estadístico) es una forma de “normalizar” los valores de una larga serie de datos. Es

una fórmula que permite la estandarización de una distribución. Se calcula restando al dato la media de la distribución y dividiendo el resultado por la desviación típica (la distancia que tiene dicho dato respecto de la media). La puntuación Z permite conocer cuán lejos está un dato concreto respecto de la media, y lo traduce de una forma muy sencilla: la media pasa a ser 0, y el dato arrojado será negativo o positivo según esté por debajo o por encima de la media. Si se quiere comparar la nota de un alumno de una asignatura con la nota de otro alumno de la misma asignatura, pero de otro profesor no sería real o acertado evaluar sin más los datos absolutos, pues los exámenes han podido ser muy diferentes. El z-score de cada alumno permite ver qué nota ha sacado cada uno de ellos en función de la media de cada clase. Es decir, z-score dirá si los alumnos que se quieren comparar están por encima o por debajo de la media de su clase y cuán lejos están. Esto permitirá comparar las notas de manera más justa, real y exacta. PF realiza un cálculo z-score para cada valor obtenido, y ello permite buscar mejor los valores atípicos (*outliers*), esenciales para la filosofía de PF a la hora de detectar patrones y comportamientos destacados en un conjunto de datos.

8. Tiempo máximo de carga de bibliotecas. Determina cuándo PF dirá que tarda demasiado en cargar una librería o biblioteca de código; ampliarlo ayuda en conexiones muy lentas a minimizar la tasa de errores.
9. Valores máximos a mostrar en estadística descriptiva. Cuántos valores máximos de campo PF exhibirá tras la consulta. PF realiza un cálculo de estadística descriptiva (valores medios, valores significativos, ...) en cada consulta. Uno de ellos es la lista de valores singulares (campo por campo) más repetidos, lo que da una pista de qué “abunda” más en una consulta. No se debe confundir esta lista con los resultados finales. Digamos que en una consulta geográfica el valor “Madrid” en el campo “localidad donde se celebra” aparece “implicado” en varias filas de la consulta (“Madrid > Museo del Prado”, “Madrid > Reina Sofía”, ...). Los valores máximos NO SON LA SUMA de las duplas “Madrid / Prado”, supongamos que 50, y “Madrid /Sofía”, supongamos que 75; la suma sería 125, pero el valor máximo de Madrid sería 2, porque aparece dos veces. Comparando ambas magnitudes se puede saber si la frecuencia de aparición está o no relacionada con su interés o significación global. Se puede hacer una consulta que detecte que Madrid, pese a estar vinculado a 175 apariciones o registros solo figura esas dos veces, lo que indicaría que su posibilidad de aparición está muy mediatizada por sus “parejas de baile”; vamos, que no es

- tan importante Madrid en sí mismo sino Madrid con los museos señalados, no con otros.
10. Límite máximo de vértices para trazar un grafo. En evaluación si se mantiene el valor en Ajustes. Indica cuándo PF disparará un aviso advirtiéndolo de un posible error de exceso de consumo de memoria al realizar el trazado de un grafo.
 11. Límite máximo de aristas para trazar un grafo. Ídem que el punto anterior.
 12. Elementos por página en consulta de registros. Cuántas filas mostrará la tabla.
 13. Enlaces a páginas anteriores y posteriores en consulta de registros. Cuántas páginas (siguientes y anteriores) se mostrarán bajo la tabla.
 14. Escala cromática. PF utiliza una gama de colores para señalar los elementos más destacados tanto en tablas, en los campos que contienen datos numéricos, como en gráficos y otras áreas. Se puede escoger cuál de ellas se prefiere. Basta con pulsar sobre su nombre (no sobre los colores). De momento figuran escalas en azules, verdes, rojos, malvas y verde a rojo (más sus inversas).

NOTA IMPORTANTE

PF ALMACENA EN UNA PARTE DE LA MEMORIA “FIJA” DEL NAVEGADOR LOS AJUSTES SELECCIONADOS POR EL USUARIO. EN OCASIONES, Y POR DIVERSAS RAZONES (EN ESPECIAL DURANTE PROCESOS DE DESARROLLO, QUE SE AVISARÁN CONVENIENTEMENTE), ES DESEABLE “LIMPIARLA” SI SE HAN PRODUCIDO CAMBIOS SIGNIFICATIVOS ESTOS VALORES. EN CASO CONTRARIO, ES POSIBLE QUE PF SE COMPORTE DE FORMA INESPERADA Y PRODUZCA ERRORES. PARA EVITARLOS REALIZAR LAS SIGUIENTES TAREAS (EJEMPLO PARA NAVEGADOR CHROME, SIMILAR, AUNQUE CON DIFERENCIAS, PARA SAFARI O FIREFOX):



Ilustración 2. “limpieza” de “Local Storage”.

1. CON PF ABIERTO, Y ANTES DE CARGAR BD, PULSAR CON BOTÓN DERECHO SOBRE UNA ZONA DESPEJADA DE LA PÁGINA (QUE NO CONTenga TEXTO O IMÁGENES).
2. EN EL MENÚ EMERGENTE, SELECCIONAR INSPECCIONAR. SE ABRIRÁ A LA DERECHA (O DEBAJO, SEGÚN CONFIGURACIÓN DEL NAVEGADOR) UNA ZONA DE CONSOLA PARA DESARROLLADORES.
3. EN LA ZONA SUPERIOR DE DICHA ÁREA, DONDE PONE ELEMENTS >>, PULSAR “>>” Y SELECCIONAR APPLICATION.
4. EN STORAGE, SELECCIONAR LOCAL STORAGE Y DESPLEGAR; APARECERÁ [HTTPS://HDPLUS.ES](https://hdplus.es).
5. PULSAR CON BOTÓN DERECHO SOBRE ESA URL Y SELECCIONAR CLEAR.

DE ESA MANERA, PF VOLVERÁ A INICIALIZAR LOS VALORES DE AJUSTES. ESTA TAREA SOLO ES CONVENIENTE REALIZARLA MIENTRAS PF ESTÉ EN PROCESO DE DESARROLLO, Y NO HAY QUE HACERLA CADA VEZ, SOLAMENTE CUANDO HAYA HABIDO ALGÚN CAMBIO SIGNIFICATIVO EN LA CONFIGURACIÓN. ESO SE SABE PORQUE EN LA OPCIÓN NUEVA NO APARECERÁ NINGÚN VALOR MARCADO.

DATOS

La pantalla de Datos concentra toda la actividad principal de la aplicación. En ella se despliegan las tablas, los gráficos, los mapas y los análisis. Los elementos principales se describen a continuación.

Panel

Con formato de árbol jerárquico horizontal, solo muestra las opciones disponibles en función de lo realizado. Cambia de forma significativa tras cargar BD, y antes de ello solo ofrece dos posibilidades: actualizar datos desde servidor EF, que indica si

el servidor está disponible y evalúa en milisegundos (ms) su velocidad de acceso, y cargar archivo. Una vez se ha cargado BD, aparecen los siguientes bloques de opciones:

1. Carga. Permite obtener una estadística global de la BD (*Más información...*) y exportar los datos filtrados en formatos TAB, CSV y XML.
2. Informes. Sobre cada tipo de registro (obra de arte, entidad, ...), y mostrando sus relaciones para determinado campo relacional (autor de obra de arte, ...) o para cualquiera. En dos formatos: analítico (compatible con hoja de cálculo) y textual (RTF, compatible con procesador de textos), éste último muy lento y trabajoso de realizar. Está en desarrollo la posibilidad de generar informes personalizados.
3. Colecciones. Son filtros preestablecidos que agrupan los registros de determinada manera imponiéndoles ciertas condiciones. Las hay predeterminadas, guardadas en el servidor PF, o calculables, guardadas en cliente y que se pueden activar desde el filtro avanzado.
4. Consultas. Filtros (ver más adelante), búsqueda (ídem) y listado.
5. Análisis. Sus opciones se corresponden con las del submenú de la página, de la 3 a la 7.

Filtro y búsqueda

PF utiliza el paradigma de *filtro*, propio de la teoría de conjuntos. Básicamente trata de seleccionar mediante inclusión o exclusión de determinadas características a un subgrupo de los registros presentes en el sistema.

Tenga en cuenta que, una vez aplicado un filtro a los datos, éste permanecerá activo hasta que no se establezca lo contrario, y su presencia afecta a TODAS las operaciones que PF ejecute.

Permite establecer condiciones sucesivas y concurrentes que acotarán el ámbito de trabajo. Para ello:

1. Establezca los *restringidores* para una condición. Tenga en cuenta que únicamente son obligatorios operador y valor.
 - a. modificador: extrae un aspecto determinado de ciertas *metainformaciones*; por ejemplo, año o día de la semana de una fecha.
 - b. metainformación: también denominado *faceta* en ciertos sistemas; determina sobre qué tipo de información se realizará la búsqueda; por ejemplo, lugar donde se celebra una exposición.
 - c. operador: determina cómo ha de cumplirse una condición; por ejemplo, que un registro contenga o no sea igual a un valor dado; tenga en cuenta que dicho

cumplimiento, si no se especifica modificador o metainformación, se verificará sobre todos los registros. Los operadores “*contenga*” y “*no contenga*” no tienen en cuenta el uso de mayúsculas, minúsculas y tildes o caracteres diacríticos. La precisión de esta característica depende del navegador empleado y puede variar, aunque no de modo significativo, entre unos y otros con signos muy poco habituales en idiomas occidentales.

- d. valor: texto a buscar (puede ser una cifra, una fecha, ...). ATENCIÓN: si utiliza el operador “*entre*” separe los valores con el símbolo “/” (barra). Es decir, si desea buscar un valor comprendido entre 2010 y 2011 escriba “2010/2011”.
2. Pulse “*calcular*” para que PF evalúe el número de registros o subconjunto que cumple la condición.
3. O pulse “*valores*” para saber qué registros cumplen la condición *antes* de evaluarla. Recuerde que la lista de valores se calculará en función del operador empleado, por lo que puede o no tener en cuenta el uso de mayúsculas, minúsculas y tildes o caracteres diacríticos.
4. Añada, si lo desea, cuantas condiciones precise para acotar el filtro. PF ÚNICAMENTE UTILIZARÁ PARA CONSTRUIR EL FILTRO AQUELLAS CONDICIONES QUE SE HAYAN CALCULADO PREVIAMENTE.

Así pues, una vez haya calculado todas las condiciones implicadas, establezca el filtro siguiendo el esquema adecuado a lo que se pretenda conseguir utilizando las siguientes opciones.

Estricto (intersección exclusiva): el filtro deberá seleccionar ÚNICA Y EXCLUSIVAMENTE los registros que cumplan SIMULTÁNEAMENTE TODAS las condiciones establecidas. Por ejemplo,

Todas (intersección): el filtro deberá seleccionar aquellos registros que cumplen todas las condiciones establecidas, aunque no necesariamente de forma simultánea.

Cualquiera (unión): el filtro deberá seleccionar aquellos registros que cumplen cualquiera de las condiciones establecidas.

Cada vez que añada, modifique o quite una condición deberá calcularla individualmente o no se incluirá en el filtro.

Dispone de una lista de modificadores, metainformaciones y operadores válidos en el apartado de *Terminología*. La ilustración que se muestra a continuación puede ayudar a comprender mejor el proceso de filtrado.

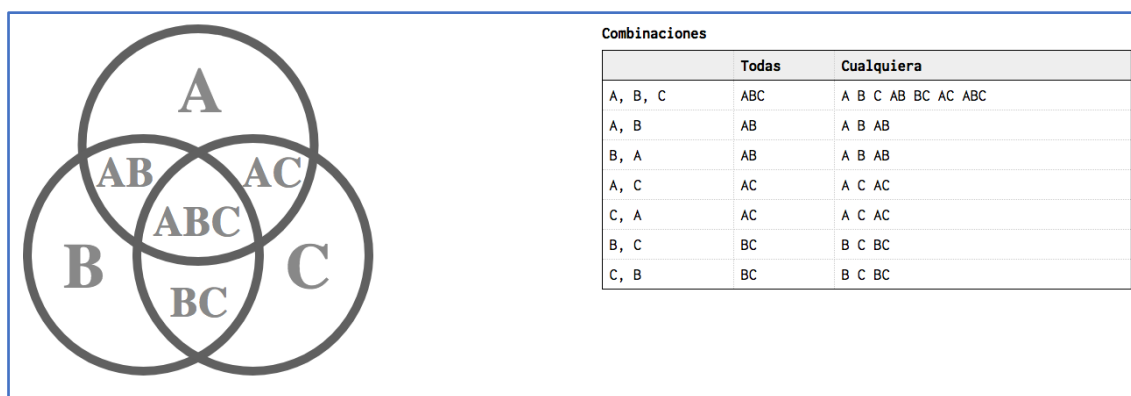


Ilustración 3. Tabla de combinaciones en el proceso de filtrado.

La búsqueda es una versión simplificada del filtro, con una única condición. Para que sea efectiva deberá seguir las siguientes convenciones.

1. Escriba todo o parte del texto que desee buscar.
2. Seleccione dónde se realizará la búsqueda: o en un tipo de metadato asociado a los registros o en cualquiera de ellos.
3. Pulse calcular para que PF evalúe los resultados obtenidos.
4. Pulse filtrar para que PF establezca el filtro. Recuerde que pulsar esta opción provocará un error si previamente no se ha realizado el cálculo.

La siguiente ilustración puede ayudar a comprender el modelo de condiciones de filtrado.

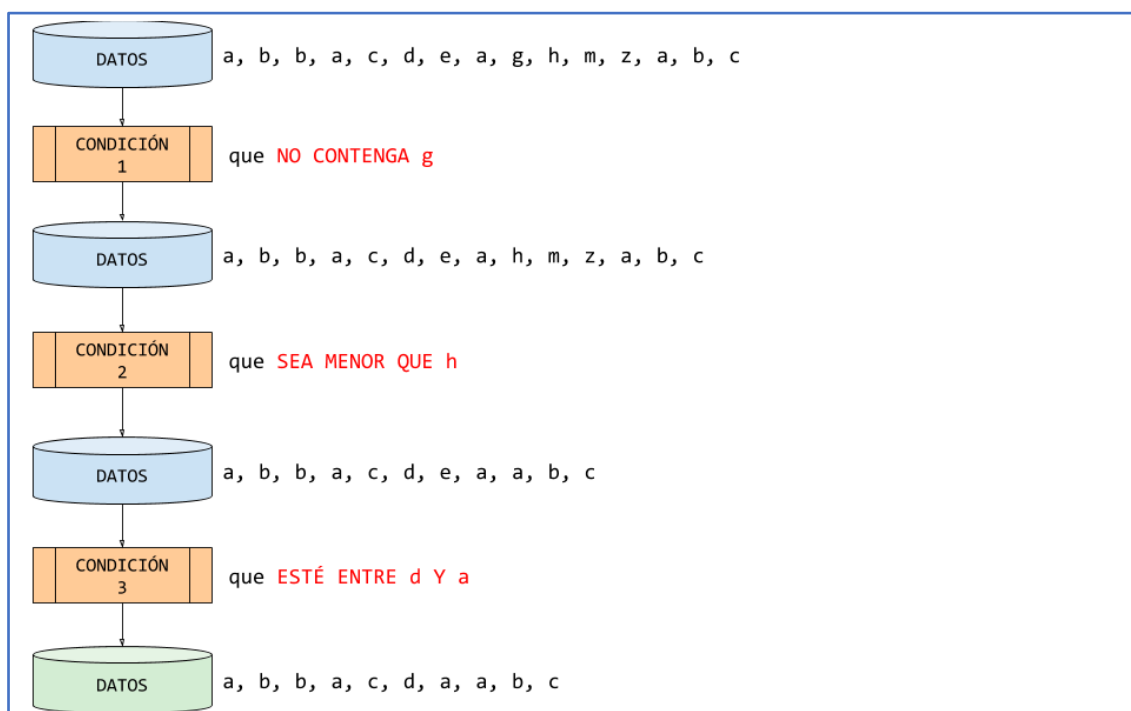


Ilustración 4. Ejemplo de condiciones de filtrado.

Listado

Muestra una visión paginada de los registros seleccionados mediante el filtro. La cantidad de filas mostradas en cada página y el número de páginas antecedentes y precedentes de la actual pueden ser configurados desde la pantalla *Ajustes*. Puede desplazarse a una página concreta. En cada fila se muestra el tipo de registro primario, su título, que es un enlace a la ventana emergente de registro individual, y sus taxonomías.

En el apartado de registro individual hay que tener en cuenta lo siguiente:

1. Al pulsar sobre un campo que muestre coordenadas aparecerá una ventana emergente con el correspondiente mapa estático y su geocodificación inversa (de coordenadas a dirección postal) calculada.
2. Cualquier elemento en formato *link* de los apartados *Taxonomía* y *Metainformación* activará una búsqueda global (no un filtro) del término enlazado.
3. Un asterisco entre corchetes ([*]) abrirá una página externa en Wikipedia sobre el término marcado.
4. Las *Relaciones* pueden ser de dos tipos: *Aparece citado por...* ...en calidad de... y *Cita a...* ...en calidad de... Todas ellas figuran agrupadas, ordenadas alfabéticamente y numeradas.
5. Cronograma. Gráfico de dispersión y líneas donde se marcan los cronovalores más importantes de los elementos relacionados con el registro. El gráfico cartesiano marca los grupos de valores con un rectángulo (por ejemplo, fecha de nacimiento para actores que participen como autores en una exposición), con una línea su media aritmética simple y con pines los valores máximo y mínimo en cada caso. Ajustable a año, mes (número o texto), día y día de la semana (número o texto).
6. Geograma. Ídem que el anterior para valores geográficos.
7. Red. Grafo de relaciones del registro principal con otros registros.
8. Taxograma. Ídem que en los casos 5 y 6 pero para las taxonomías.
9. Estadísticas. Tabla pivote para recuento por campos y valores para el registro principal y sus relaciones.
10. Mapas. Necesita conexión a la red. Puede mostrar el registro principal, los agrupamientos (*clusters*) de sus relacionados y de los relacionados con sus relacionados o “relaciones indirectas”, el mapa de calor de ambos grupos de valores y los vínculos correspondientes en forma lineal.

Criba

Permite trabajar sobre el subconjunto de registros filtrados y realizar recuentos mediante su agrupación por hasta seis niveles de anidamiento. Sus valores son graficables (ver más adelante). Tiene dos formas de mostrarse: *tabla de resultados* y *estadística*.

Antes de seleccionar permite marcar las condiciones de agrupación: estricto (solamente se seleccionarán filas cuyos campos contengan algún valor en todos los casos) y solo significativos (solamente se seleccionarán los registros con la condición de anomalía, los denominados *outliers*).

Desde ese mismo párrafo de filtrado se puede escoger entre las vistas de tabla y estadística, y exportar los datos. Para cribar se seleccionará un tipo de registro (obra de arte, entidad, ...) y hasta 6 subniveles de agrupamiento, incluyendo campos relacionales, con sus modificadores correspondientes (por ejemplo, “autoría de la obra ID”, “autoría de la obra título” o “autoría de la obra texto”, correspondiendo, respectivamente, a “NNNN”, “Pablo Ruiz Picasso” y “NNNN: Pablo Ruiz Picasso”). Así es posible realizar consultas del tipo “exposiciones por década de fecha de apertura, región y localidad agrupadas por entidad organizadora, actor autor y temática”. Una vez realizado el recuento, PF mostrará tres campos adicionales:

1. “normalizado”. Valor de *z-score* reducido a +/- 5, en números o estrellas.
2. “significativo”. Si el resultado entra o no en el grupo de *outliers*.
3. “recuento”. Valor absoluto. Coloreado mediante escala cromática (ver *Ajustes*).

Bajo la lista de selectores se mostrará la condición de agrupamiento, las filas obtenidas, la suma total de recuento, el promedio (total / filas) y los valores significativos (número de ellos y filas que forman parte). Es posible también obtener un gráfico de líneas *ordenado* (tal como figuran las filas en la tabla) o *normalizado* para observar hasta qué punto la distribución de valores es o no “normal” (función gaussiana). En este último formato el gráfico debe dibujar una “campana” cuya cima debe aproximarse al centro del gráfico; cuanto más se desvíe del centro más sesgado estará el conjunto hacia los valores menores o mayores. Supongamos que agrupamos actores por edad en años, género y país de nacimiento. Si la campana se aproxima más o menos a uno de los dos extremos indicará el sesgo de la muestra.

El formato *Estadística* realiza una serie de operaciones de cuantificación del subconjunto de datos que arrojan luz sobre la

distribución de los mismos y generan nuevo conocimiento. Básicamente se realizan una serie de cálculos encuadrables en el ámbito de la estadística descriptiva, evaluando parámetros como las medidas de centralización y de dispersión. Consta de 4 zonas.

Muestra

Computa un histograma de los datos. Básicamente es una representación numérica de la frecuencia de aparición de los valores agrupándolos en grupos (*bins* o *buckets*, según terminología); algo así como cuando se representan rangos de edades: “de cero a cuatro, de cinco a nueve, ...”. Sirve para averiguar qué grupo de valores es el más numeroso, es decir dónde se concentra la significación. Hay que tener cuidado: marca agrupamientos en los recuentos, no en los valores textuales. Es decir, si indica que el 98% de los valores (1, 3, 9, ...) está en el grupo de 0 a 24, y que el 2% se encuentra en el grupo de 2.000 a 2.024 significará que los datos textuales asociados a un valor que no supere las 24 apariciones será lo más frecuente, mientras que un valor con un recuento de 2.010 será raro e infrecuente.

El histograma se usa para elaborar una distribución de Pareto y ver hasta qué punto los valores seleccionados se asemejan a lo previsto. Vilfredo Pareto, a finales del XIX, afirmó que en un conjunto cualquiera de resultados normalmente había “muchos triviales y unos pocos vitales”; se la conoce como la regla de 80-20 (80% del montón, 20% significativos), análisis A-B-C, y otras denominaciones. PF traza dicho diagrama.

El apartado *Muestra* también entrega una serie de cálculos interesantes situados en cabecera:

1. Total. Cuenta del conjunto de valores.
2. Único. Cuenta de valores diferentes (si la cifra “1” aparece dos veces solo contará una vez).
3. Moda. Lista de valores más repetidos, separada por comas.
4. Desviación media. La desviación media absoluta (DMA o MAD por sus siglas en inglés) de un conjunto de datos es la distancia media entre cada valor y el promedio. La desviación media absoluta es una manera de describir la variación en un conjunto de datos. Es la “media de la distancia a la media”, es decir, cuánto se separan los valores de lo que sería su media aritmética.
5. Rango. Número de valores diferentes entre el máximo y el mínimo de la muestra.
6. Entropía. Es un concepto discutido en el ámbito de la estadística, y proviene de la disciplina de la Física, aunque se utiliza con asiduidad en Telecomunicaciones. Mide en grado de similitud entre los elementos que integran un conjunto de datos y si entre ellos los

- valores se repiten con frecuencia o no; en gran medida, caracteriza la predictibilidad de una muestra. PF entrega el valor absoluto sobre su máximo posible (4,75/6, por ejemplo). Es un concepto novedoso en el ámbito de las Humanidades Digitales tal y como PF lo utiliza.
7. Homogeneidad. Mide la parte “no entrópica” de la muestra. Es decir, si la entropía es alta (como en el ejemplo anterior) significará que la homogeneidad de la muestra (hasta qué punto unos valores se parecen a otros) es baja.
 8. Valores significativos. Cuenta cuántos *outliers* figuran en la muestra. El concepto de anomalía estadística es muy importante en el ámbito de la HAD, porque “señala con el dedo”, por así decirlo, a todo aquello que “se sale de la normalidad”. El método matemático más generalizado para detectarlos se conoce como test de Tukey, y utiliza la diferencia entre el primer y el tercer cuartil, que suele denominarse rango intercuartílico o IQR (*interquartile range*). Para calcular estos elementos en primer lugar se localiza la mediana de la muestra, es decir, el valor de la variable situada en la posición central de un conjunto de datos ordenados. A partir de ahí se ubican los cuartiles, que son los tres valores que dividen a dicho conjunto en cuatro partes porcentualmente iguales; es decir, que donde comienza el 25% de los elementos se sitúa el primer cuartil, la mediana en el 50% y el tercer cuartil en el 75%. Suelen considerarse dos categorías de valores atípicos:
 - a) Leve ($1.5 * IQR$): Si el valor está a 1,5 veces la distancia del rango intercuartílico con respecto a uno de esos cuartiles.
 - b) Extremo ($3 * IQR$): Si el valor está a 3 veces la distancia del rango intercuartílico con respecto a uno de esos cuartiles.
 Los valores atípicos, por tanto, son aquellos que se sitúan “por debajo” o “por encima” de la “normalidad”. PATHFINDER, por razones de operatividad, detecta valores atípicos de categoría leve ($1.5 * IQR$).
 9. Mediana, primer y tercer cuartiles. Para complementar la información del punto anterior.

Campos

Tabla con cada “campo” o tipo de valor individualizado y los cálculos descriptivos correspondientes, además de su entropía. Los valores mostrados son los siguientes: FLD = campo; VLD = si su valor válido (no es nulo); DST = valores únicos; MIS = valores ausentes; MEA = media aritmética; MED = mediana; STD = desviación típica; MSK = asimetría de moda (hasta qué punto la moda, es decir, los valores más repetidos, se acerca al extremo

más bajo o más alto del conjunto de datos); Q1 = primer cuartil; Q3 = tercer cuartil

Afinidad

En desarrollo. Trata de computar cuán relacionados están entre sí los valores singulares de los campos que se aluden en el punto anterior. Es decir, intenta averiguar que el hecho, por ejemplo, de que, en una criba por países de origen de los actores y su género, en el caso de que aparezcan muchas veces “España” (recuérdese que hablamos de cuenta del número de veces que un valor aparece, no de la suma de sus recuentos en combinación con otros) y “femenino” está o no relacionado y puede detectarse (o no) una *causalidad*. Utiliza un concepto familiar para las DH en su versión de análisis textual, el índice de Jaccard, que mide el grado de similitud entre dos conjuntos, sea cual sea el tipo de elementos. Es decir, la cardinalidad de la intersección de ambos conjuntos dividida por la cardinalidad de su unión. En pocas palabras, se calcula cuántos elementos tienen en común dos conjuntos de números y se lo divide entre cuántos tienen en total ambos.

Esta medida solo estará presente si de los 6 posibles niveles de agrupamiento al menos se han completado 2, y empezará a tener algo de significación con 4 ó más de ellos.

Relaciones

Muy similar a la criba, pero partiendo de un concepto diferente: recuenta relaciones entre registros. Este apartado es el que está más relacionado con la carga de la BD y el punto correspondiente en *Ajustes*. El modelo de recuento es simple:

[tipo] → [campo de dicho tipo con o sin modificador] → [tipo de relación: cita a o es citado por] → [campo del registro relacionado con o sin modificador] → clave (tipo de relación; en PF se muestra con una traducción errónea como “registro”)

Los elementos entre corchetes son seleccionables por el usuario, y el último es calculado. Así se pueden realizar consultas que responden, por ejemplo, a “¿cuántas veces Picasso aparece citado como autor por exposiciones agrupándolo por años?”.

Los procesos de tabla y estadística funcionan igual que en Criba.

Gráficos

Frank Anscombe, profesor de Cambridge, allá por 1973, publicó un interesante estudio titulado *Graphs in Statistical Analysis*, en el que enfatizaba sobre la necesidad de “mirar” los datos antes

de analizarlos. Para ejemplificar el por qué de tal afirmación propuso un conjunto de datos estadísticamente muy similares y los mostró en cuatro gráficos diferentes, el célebre “cuarteto de Anscombe”; al hacerlo transmitía perfectamente la idea: ni se parecían, mientras que los valores apenas diferían entre sí.

Esa premisa forma parte de PF y de las ideas esenciales de su desarrollo desde el principio, y está presente, por ejemplo, en los gráficos que aparecen en los registros individuales de *Listado*.

Este submenú, con alguna opción aún bajo proceso de construcción pero ya plenamente funcional, tiene ese propósito. Para ello ofrece (de momento) gráficos de líneas con ciertas particularidades tanto para *consultas específicas* como para los resultados de *Criba* y *Relaciones*.

Es importante comprender bien el funcionamiento de los diferentes elementos del interfaz gráfico para aprovechar todas sus posibilidades.

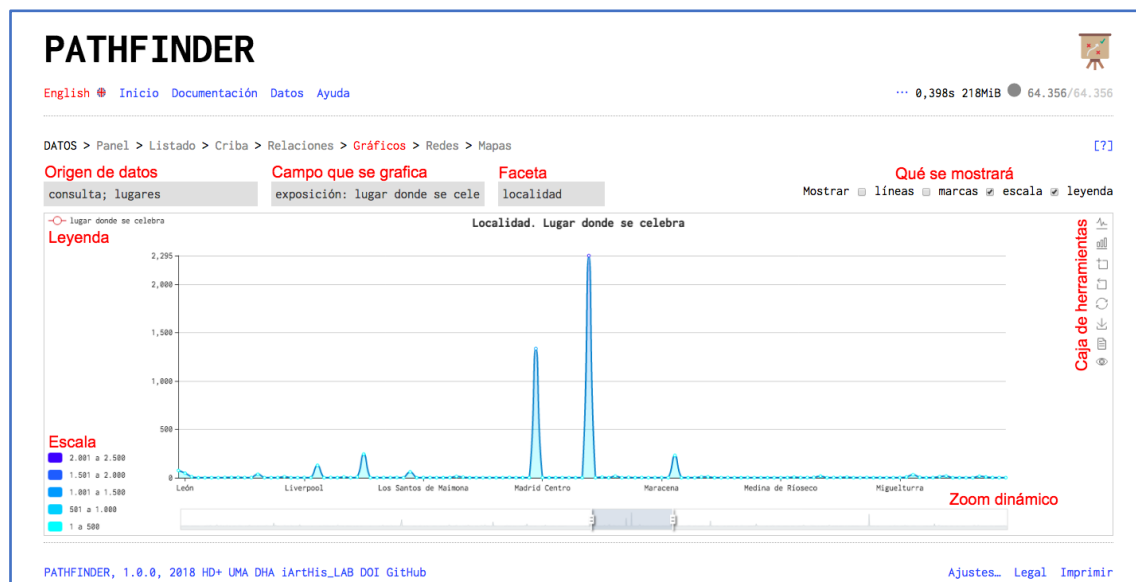


Ilustración 5. Gráficos.

En la zona de *selectores* figuran 3 elementos:

- Origen de datos. Subconjunto de valores a graficar. De tres posibles tipos:
 - a. Consulta. Específicas por temas de interés. Afectan a la totalidad de datos de la BD filtrados. Por eso es importantísimo filtrar previamente los datos a considerar. Pueden ser:
 - i. Lugares. ¿Dónde se celebran las exposiciones o nacen los actores...?
 - ii. Edades o duraciones. ¿Qué edad tienen los actores o cuánto duran las exposiciones o la creación de una obra...?

- iii. Géneros. Específico para autores.
- iv. Servidores de Internet. ¿Qué servidores se emplean para difundir los datos en formato HTML o RSS?
- v. Fechas de comienzo. ¿Cuándo empieza la exposición, se crea la obra o nace el actor?
- vi. Fechas de finalización. Ídem que el anterior, pero para finales.
- vii. Taxonomías.
- b. Tabla. Datos de las tablas de Criba o Relaciones.
- c. Estadística (aún no implementado, pero ya en desarrollo). Ídem que el anterior, pero afectando a la estadística descriptiva. Probablemente sea un gráfico de dispersión con agrupamiento en clusters.
- Campo que se grafica. Delimita el tipo de información a graficar en función del conjunto de datos. Por ejemplo, para edades o duraciones serían obra de arte: tiempo de ejecución de la obra, catálogo: antigüedad desde publicación, actor: edad y exposición: duración. Es posible mostrar simultáneamente todos los campos.
- Faceta. Modificador para el campo seleccionado. Por ejemplo, en el caso anterior, días, meses, semanas y años.

En *Qué se mostrará* se puede elegir si el gráfico muestra:

- Líneas. Mediana, primer cuartil, tercer cuartil.
- Marcas. Máximo, mínimo y *outliers*. Los dos primeros cambian si se utiliza el zoom dinámico para ajustar el rango de visión.
- Escala: distribuye los resultados en histograma; al pulsar sobre uno de los *buckets* muestra u oculta los valores correspondientes.
- Leyenda. Al pulsar sobre el símbolo muestra u oculta los valores correspondientes.

Tanto la escala como el área del gráfico (con líneas suavizadas) se adecua a la escala cromática elegida en *Ajustes*. El color de la línea se corresponde con el asignado a cada tipo de registro.

En la barra de herramientas hay algunos elementos destacados. De arriba abajo son los siguientes.

- Alternar entre gráfico de líneas y gráfico de barras.
- Selección gráfica (con ratón) del área a ampliar y eliminación de la selección.
- Restaurar. Usado para cuando se alterna entre líneas y barras, para reponer el formato original.
- Descarga de la imagen del gráfico en formato PNG.
- Tabla de datos. Permite ver y exportar los datos subyacentes en el gráfico en forma de tabla.

- Alternar entre pantalla completa y pantalla normal, ajustada.

Quizá lo más interesante sea el zoom dinámico. Muestra una miniatura del gráfico para la totalidad de los datos considerados y dos “manijas” que permiten reducir o ampliar lo que se muestra en el gráfico superior. Facilita la observación detallada.

Redes

Una advertencia preliminar: es MÁS QUE CONVENIENTE solo realizar grafos de redes UNA VEZ SE HAYA ESTABLECIDO UN FILTRO. Por poner un ejemplo concreto, con 65.536 registros, PF calcula 277.778 relaciones, y el grafo contiene 57.466 vértices, 138.894 aristas, 19 claves de tipos de arista y 6 categorías de vértices, ello solo considerando la relación “aparece citado por...”. Un cliente con +16GiB RAM puede con ello, a uno de +8GiB le cuesta, y uno de +4GiB tiene muchísimas dificultades. No es una limitación de PF, sino meramente matemática. Los cálculos de grafos nunca son sencillos, sobre todo cuando se realizan en un “dibujo” en pantalla (se pueden hacer solamente con ecuaciones).

Este submenú está aún en desarrollo activo, e irá incorporando las novedades pertinentes.

Hay cuatro tipos de presentación del grafo (*layouts*), si bien el último probablemente será reemplazado por otras variedades sobre las que el desarrollo está incidiendo (diagrama de matriz o coocurrencias, diagrama de colmena o *hive chart* y diagrama de acordes o *chord diagram*). En ellos, el tamaño de los vértices se corresponde con su peso proporcional bien ajustado; es decir, si en el conjunto estudiado Picasso aparece relacionado con 200 elementos ese será su tamaño; el mínimo que se graficará es de 10. El color también se corresponde con el tipo de registro. Al situar el cursor sobre un vértice se resaltan sus aristas y los vértices relacionados, además de mostrarse su valor en un *tooltip* informativo. Los tipos son:

- Simple. Aristas unidas mediante vértices curvados. Es ampliable pero los vértices no se pueden desplazar individualmente.
- De fuerza. “Sometido” a atracción y repulsión “gravitatoria”. Los vértices se pueden desplazar individualmente y es ampliable.
- Circular. En “corro”, por así decirlo. Ampliable pero no desplazable individualmente.
- Sankey. Representa un tipo específico de diagrama de flujo, en el que la anchura de las flechas se muestra proporcional a la cantidad de flujo, y se usó por primera

vez en 1898 en una publicación sobre la eficiencia energética de la máquina de vapor. Representa bien cuál es la importancia de cada relación (es decir, mejora la percepción de las aristas y no de los vértices). Desgraciadamente no siempre es representable (necesita ciertas condiciones muy concretas), por lo que será sustituido por versiones más eficaces.

Mapas

En pleno desarrollo. Básicamente mimetizan lo que se muestra en la ficha individual en *Listado* pero con ciertas condiciones.

La información sobre el tema se añadirá posteriormente al presente documento.

OTRAS CONSIDERACIONES

Todas las tablas son “ordenables” al pulsar alternativamente sobre el nombre de una columna (orden ascendente y descendente), con la excepción de Listado. Es así por diseño, y aunque es posible modificarlo de momento sería una más que notable pérdida de tiempo en detrimento de otras áreas de desarrollo más urgentes.

Aunque se ha abundado en el tema del filtro vale la pena hacer unas observaciones finales.

En primer lugar, hay que recordar que SIEMPRE QUE SE EMPLEE EL OPERADOR “CONTenga” o “NO CONTenga” no se consideran mayúsculas, minúsculas y tildes; la ausencia de consideración de los diacríticos depende de la implementación de Javascript (lenguaje de programación empleado en PF) que cada navegador tenga. El aplicativo está probado al respecto con éxito en Safari, Chrome y Firefox; otras implementaciones pueden diferir en sus cálculos.

En segundo lugar, el resultado se obtendrá a partir del vínculo entre las condiciones de filtro, cada una de ellas singular en sí misma. Ello se realiza de acuerdo a lo expresado en la siguiente tabla.

Estricto	Todo	Cualquiera
Sólo lo que comparten los conjuntos obtenidos a partir de las condiciones A, B y C; es decir, que X sea parte de A, sea parte de B y sea parte de C	Cualquier cosa que compartan A, B y C; es decir, si X pertenece a A y B pero no a C, se incluirá; igualmente si pertenece a C y B pero no a A. No se incluirá si no pertenece <i>a la vez</i> a una pareja de condiciones al menos	Cualquier cosa que pertenezca a cualquiera de las condiciones A, B o C, tanto si está compartido como si no

Hay que considerar las condiciones de filtro en su conjunto para obtener resultados. Si para saber en cuántas exposiciones Picasso participó como autor en Málaga entre 2010 y 2015 se utiliza el siguiente juego de condiciones y se emplea Estricto o Todo para calcular el resultado el conjunto obtenido sería incorrecto (en realidad sería “cero registros”).

1. QUE [SEDE] CONTENGA "REINA SOFIA"
2. QUE [ACTOR] CONTENGA "PICASSO"

¿Por qué? Porque PF “piensa” lo siguiente: “este usuario quiere saber cuántas exposiciones se han celebrado en el Reina Sofía y cuántos actores se llaman ‘ruiz picasso’; pero ¿qué tienen que ver las exposiciones con los actores? Como no lo sé porque no me lo ha dicho, no puedo obtener nada”.

Y no deja de ser lógico. Sin embargo, si se realiza la consulta con los siguientes parámetros, la cosa cambia:

1. QUE [SEDE] CONTENGA "REINA SOFIA"
2. QUE [ACTOR] CONTENGA "PICASSO"
3. QUE [ACTOR QUE PARTICIPA COMO ARTISTA] NO SEA NULO

Con BD a fecha 20180821, PF arroja 993 registros si el vínculo entre condiciones es “Todos”. Porque ya sabe que queremos averiguar en cuántas exposiciones del Reina Sofía Picasso participó en calidad de artista, mientras que antes no era posible deducirlo.

Ahora bien, si solamente se desea saber qué exposiciones y no qué artistas, en el apartado *Incluir* del filtro avanzado solamente dejaremos marcada su casilla.

Es preciso recordar que, al lado de cada condición, hay un vínculo que pone “valores”. Mostrará una ventana emergente con los posibles valores que el campo podrá adoptar, pero siempre teniendo en cuenta el operador seleccionado; es decir, si pulsamos “valores” tras haber escrito “ruiz picasso” en el campo de la condición y con el operador “sea igual que” seleccionado no nos mostrará nada, puesto que no hay ningún actor cuyo nombre sea exactamente igual que “ruiz picasso”; sin embargo, sí mostrará alternativas si el operador es “contenga”.