El Manual de la OMPI sobre análisis de patentes de código abierto

Tabla de contenido

Portada

0.1 autoría

El Manual fue escrito por el Dr. Paul Oldham de One World Analytics , con contribuciones de y bajo la coordinación de la Sra. Irene Kitsara (OMPI). **2016**

0.2 Agradecimientos

0.3 Información adicional

La versión electrónica del informe, así como todos los conjuntos de datos mencionados en el Manual y utilizados para los diversos ejercicios se pueden descargar desde el repositorio de Github Manual.

0.4 Condiciones de uso

Le invitamos a utilizar la información proporcionada en esta publicación, pero cite la OMPI y el Manual como la fuente. Correo electrónico de contacto : patent.information@wipo.int

0.5 descargo de responsabilidad

Este manual no constituye una presentación exhaustiva de todas las herramientas de código abierto y la información contenida en el mismo era válida en el momento de la impresión. Además, las opiniones expresadas en el Manual no reflejan necesariamente la opinión de los Estados miembros de la OMPI.

Capítulo 1 Introducción

Este libro proporciona una guía práctica de herramientas de software de código abierto y gratuito para el análisis de patentes. El objetivo del Manual de la OMPI sobre análisis de patentes de código abierto es proporcionar una introducción práctica al análisis de patentes sin asumir el conocimiento previo de patentes o lenguajes de programación.

Análisis de patentes de código abierto

Una característica de las herramientas de software libre y de código abierto es que esta área está avanzando rápidamente. En respuesta a esto, el Manual se divide en dos versiones:

- La versión electrónica del Manual que puede actualizarse a medida que se actualicen las herramientas.
- Un manual de referencia impreso que proporciona una guía de herramientas básicas.

El Manual se basa en la experiencia generada en el desarrollo de los Patent landscape reports (PLRs) de la OMPI en una amplia gama de temas que sirven como obras de referencia clave para los métodos de análisis de patentes. El Manual está dirigido principalmente a investigadores, profesionales de patentes y oficinas de patentes en países en desarrollo. Sin embargo, esperamos que sea de mayor interés para los investigadores y profesionales de patentes.

Los datos de patentes son importantes porque son una fuente valiosa de información técnica que puede informar a la toma de decisiones sobre si seguir o no una vía particular de investigación y desarrollo, si se debe otorgar una licencia a una tecnología particular o si se debe perseguir el desarrollo de productos en mercados particulares. Los datos de patentes también son importantes en términos económicos y de políticas porque proporcionan un indicador clave y una visión de las tendencias en ciencia y tecnología. Los datos de patentes son utilizados comúnmente por organizaciones como la OCDE , EUROSTAT y otras para informar sobre las tendencias en investigación y desarrollo. Los investigadores utilizan cada vez más los datos de patentes para investigar áreas nuevas y emergentes de la ciencia y la tecnología, como la edición del genoma o las tecnologías de adaptación al cambio climático.

La actividad de las patentes también puede ser controvertida. Las controversias importantes en los últimos 20 años incluyen patentes de ADN, patentes de software, patentes sobre métodos de negocios, el aumento de los 'trolls' de patentes y las implicaciones de la internacionalización de la actividad de patentes para los países en desarrollo. **El software libre y los movimientos de código abierto (basados en las flexibilidades de la ley de derechos de autor) son en parte una respuesta a las controversias que surgieron en torno a modelos de software propietarios que involucran derechos de autor y patentes y un deseo de hacer las cosas de manera diferente**. **Esto ha llevado a nuevos modelos para compartir datos, cooperación en innovación y nuevos modelos de negocios. En particular, una amplia gama de herramientas de software libre y de código abierto ahora están disponibles para investigación y análisis.**

1.1 Estructura

Nos centraremos en responder **dos preguntas principales**:

1) ¿**Cómo obtener datos de patentes** en un formulario que sea útil para diferentes tipos de análisis?
2) ¿**Cómo ordenar, analizar, visualizar y compartir datos de patentes utilizando código abierto y software libre**?

Al abordar estos temas, organizaremos el Manual y los materiales en **cinco temas principales:**

1) Una descripción general de las herramientas de **código abierto y software libre**
2) Acercándose a los **datos de patentes**
3) **Obtención de datos de patentes**
4) **Limpieza y puesta en orden de los datos de patentes**
5) **Análisis y visualización de datos de patentes**.

Como un proyecto centrado en herramientas de código abierto y gratuitas, todos los datos y herramientas desarrollados para el manual están disponibles a través del repositorio de proyectos de GitHub . Te animamos a que eches un vistazo al repositorio. Para comenzar con GitHub y descargar todos los materiales del Manual, instale GitHub y luego clone el repositorio. En realidad es mucho más fácil de lo que parece.

Ahora veremos rápidamente los antecedentes de los temas.

1.1.1 Una descripción general de las herramientas de software libre y de código abierto

Comenzamos el Manual con un capítulo de Información general que revisa el número cada vez mayor de herramientas de software libre y de código abierto que están disponibles para diferentes pasos en el proceso de análisis de patentes. La gran cantidad de herramientas relevantes es casi abrumadora y una característica de las herramientas de código abierto es que todas requieren inversiones de tiempo valioso para aprender cómo funcionan. En algunos casos, esto puede requerir adquirir habilidades de programación. **Para ayudar en la toma de decisiones sobre si invertir o no en una herramienta en particular, concluimos el Resumen con una lista de 12 preguntas que tal vez desee considerar**. **La más importante de estas preguntas**, y el principio que guía nuestra selección de herramientas para el Manual, **es: ¿Funciona para mí?**

1.1.2 Acercándose a los datos de patentes

Al preparar el Manual, no asumimos ningún conocimiento previo del sistema de patentes ni de las herramientas de código abierto. Para ayudarlo a comenzar, un capítulo sobre campos de datos de patentes proporcionamos una breve introducción a la estructura de los documentos de patentes y los principales campos de datos que se utilizan en el análisis de patentes.

1.1.3 Obtención de datos de patentes

Un gran desafío para comprender las implicaciones de la actividad de patentes, ya sea en campos como las tecnologías de cambio climático, software o productos farmacéuticos, **es acceder y comprender los datos de patentes.**

Los últimos años han sido testigos de un importante cambio hacia el uso de herramientas de investigación de código abierto y la promoción del acceso abierto a datos científicos junto con la promoción de la ciencia abierta. **Uno de los propósitos principales del sistema de patentes es hacer que la información sobre invenciones esté disponible para un uso público más amplio. El sistema de patentes ha respondido a esto mediante la creación de bases de datos de acceso público, como la base de datos espacenet de la Oficina Europea de Patentes, que contiene millones de registros de patentes de más de 90 países y organizaciones. OMPI Patentscope** , proporciona acceso a 52 millones de documentos de patentes y publicaciones semanales de solicitudes del Tratado de Cooperación en materia de Patentes. **Otras iniciativas para hacer que los datos de patentes estén disponibles incluyen Google Patentsy The Lens y patentes gratis en línea** . La mayoría de estas herramientas no requieren conocimientos de programación. Sin embargo, **los Servicios de Patentes Abiertas de la Oficina Europea de Patentes ofrecen acceso gratuito a los datos de patentes sin procesar para aquellos que deseen trabajar utilizando una Interfaz de Programación de Aplicaciones (API) y para analizar los datos sin procesar de XML o JSON.**

En el caso de los Estados Unidos, **es posible descargar de forma masiva toda la colección de la USPTO a través de la descarga masiva de Google de las patentes de la USPTO** . La USPTO también ha adoptado recientemente datos abiertos a través de la creación de un nuevo portal de datos y la base de datos de búsqueda de Patentsview y la API JSON . Una gama de proveedores comerciales como Thomson Innovation y PatBase , entre otros, brindan acceso a datos de patentes y, en el caso de Thomson Innovation, agregan información adicional a través del Índice de Patentes Mundial de Derwent . Como tal, existe un ecosistema de fuentes de información de patentes y proveedores.

Como veremos, el problema clave que enfrentan los analistas de patentes que usan herramientas gratuitas es obtener datos de patentes en la cantidad y con

la cobertura necesaria, y con los campos deseados para fines analíticos. El Manual recorrerá los diferentes servicios de información y explicará en detalle los servicios gratuitos que son más útiles para el análisis de patentes.

1.1.4 Limpieza y puesta en orden de los datos de patentes

Cualquier persona familiarizada con el trabajo con datos sabrá que **la mayoría del trabajo se realiza con datos de limpieza antes que el del análisis**. En particular**, los datos de diferentes bases de datos de patentes generalmente involucran diferentes desafíos de limpieza**. La mayoría de estos desafíos **implican limpiar los nombres de los inventores y los solicitantes o limpiar los campos de texto antes del análisis.**

Dos capítulos principales en el Manual tratan problemas de limpieza de datos. El primero es un capítulo **en Open Refine (anteriormente conocido como Google Refine**) que explica el proceso de limpieza de los nombres de los solicitantes e inventores para un conjunto de datos de muestra. El segundo capítulo se centra en **el uso de R para ordenar los datos de patentes en una infografía**.

Al trabajar con el Manual, le sugerimos que encuentre útiles los siguientes recursos. La primera aborda la cuestión de cómo prepararse mejor para el trabajo en análisis y la segunda aborda los problemas clave en el formato de los datos que informan el trabajo en el Manual usando R y RStudio (entorno de desarrollo integrado (IDE) para el lenguaje de programación R, dedicado a la computación estadística y gráficos).

- The Elements of Data Analytic Style de Jeff Leek (disponible de forma gratuita si es necesario)
- Hadley Wickham en Tidy Data y este video

Le sugerimos que eche un vistazo a estos documentos porque contienen ideas centrales para enfoques efectivos para trabajar con datos de patentes.

1.1.5 Análisis y visualización de datos de patentes

Las preguntas centrales en el análisis de patentes son: ¿quién, qué, dónde, cuándo, cómo y con qué? La forma en que abordemos estas preguntas dependerá del objetivo del análisis de patentes. Sin embargo, en casi todas las circunstancias, darse cuenta de que el objetivo dependerá de las combinaciones de respuestas a las preguntas básicas. La visualización de datos de patentes es una característica esencial del análisis de patentes moderno. En pocas palabras, los humanos absorben mejor la información visual que las columnas y filas de números o grandes números de textos.

Dos capítulos principales en el Manual abordan la visualización de datos de patentes utilizando tableros con Tableau Public y gráficos interactivos usando Plotly con archivos de Excel o usando RStudio . La visualización de redes de solicitantes, inventores o tecnologías es una característica creciente del análisis de patentes y proporcionamos un recorrido práctico utilizando el software de código abierto Gephi. Con la creciente popularidad de la infografía, también se proporciona un capítulo central sobre la preparación de datos para una infografía utilizando RStudio y el servicio de infografía en línea infogr.am .

Mirando más allá del análisis y la visualización de patentes, en el Manual principal incluimos un capítulo sobre cómo se puede usar RStudio para acceder a la literatura científica usando paquetes desarrollados por ropensci para acceder a la Biblioteca Pública de Ciencias rplos como una introducción para acceder a la literatura científica más amplia usando paquetes como como full text.

1.1.6 Compartir datos y la redacción del manual

Al escribir este Manual, decidimos en una etapa temprana utilizar herramientas de código abierto y gratuitas. Nuestra herramienta de elección fue RStudio porque nos permitió escribir el Manual en markdown (rmarkdown), incluidas las imágenes y los gráficos generados a partir del código, y luego exportar fácilmente los resultados a Word, .pdf y html. También pudimos crear fácilmente un hogar para el Manual en Github y usar jekyll para lanzar versiones anteriores de capítulos como artículos tal como fueron escritos. Cuando movimos el Manual a su versión final, pudimos aprovechar el nuevo paquete de Bookdown en la versión preliminar de mediados de 2016 de RStudio para convertir el manual en el libro electrónico que estás leyendo. Todo esto fue gratis. El único requisito era la inversión adquiriendo los conocimientos para utilizar las herramientas.

Un objetivo clave detrás del desarrollo del Manual también fue poner a disposición una serie de conjuntos de datos de patentes reales que los lectores podrían usar para experimentar con las diferentes herramientas y seguir el Manual como una guía práctica. Github demostró ser ideal para esto, particularmente con la introducción del almacenamiento de archivos grandes. Si bien estas herramientas inicialmente no eran familiares e incluían una curva de aprendizaje, el proceso resultó tan fácil que **todo el Manual se escribió en rmarkdown dentro de RStudio y se publicó en el sitio web de desarrollo del proyecto en Github tal como estaba escrito.**

Esta combinación de herramientas demostró ser una forma poderosa y altamente flexible de compartir datos en bruto, resultados y análisis de una manera transparente y de fácil acceso para una variedad de audiencias. Además todas las herramientas son gratuitas. Si bien este enfoque no se adaptará a situaciones en las que la confidencialidad sea una preocupación clave, para los proyectos en los que

se pretende que los resultados sean públicos, esta combinación de herramientas representa una solución poderosa y refrescante para el antiguo problema de cómo hacer que los resultados de la investigación estén disponibles al máximo posible audiencia de forma gratuita.

Capítulo 2 Una visión general de las herramientas

Este capítulo proporciona una descripción general de las herramientas de software libre y de código abierto que están disponibles para el análisis de patentes. El objetivo del capítulo es servir como una guía de referencia rápida para algunas de las herramientas principales del kit de herramientas. Vamos a profundizar en algunas de estas herramientas en otro lugar del Manual de la OMPI sobre análisis de patentes de código abierto y le permitiremos explorar el resto de las herramientas por sí mismo.

Antes de comenzar, es importante tener en cuenta que cubrimos solo una fracción de las herramientas disponibles que están disponibles. Simplemente hemos tratado de identificar algunas de las herramientas más accesibles y útiles. **La minería de datos y la visualización están creciendo rápidamente hasta el punto de que es fácil sentirse abrumado por la variedad de opciones**. La buena noticia es que hay algunas herramientas gratuitas y de código abierto de muy alta calidad. La dificultad radica en **identificar aquellos que satisfagan mejor sus necesidades específicas en relación con sus antecedentes y el tiempo disponible para adquirir algunas habilidades de programación**. Esa decisión dependerá de usted. Sin embargo, para evitar la frustración, será importante reconocer que las diferentes herramientas toman tiempo para dominarlas. En algunos casos, como R y Python, existen muchos recursos gratuitos para ayudarlo a dar los primeros pasos en la programación. Al tomar una decisión sobre una herramienta para usar, piense detenidamente en el nivel de soporte que ya existe. Intente utilizar una herramienta con una comunidad de usuarios activa y preferiblemente grande. De esa manera, cuando te quedas estancado, habrá alguien por ahí que se haya topado con problemas similares que podrán ayudarte. Sitios como El desbordamiento de pila es excelente para encontrar soluciones a problemas.

Este capítulo está dividido en 8 secciones:

1) Herramientas generales
2) Herramientas de limpieza
3) Minería de datos
4) Visualización de datos
5) Visualización en red
6) Infografia
7) Mapeo geográfico
8) Extracción de textos

En algunos casos, las herramientas son multifuncionales y, por lo tanto, pueden aparecer en una sección donde también pueden aparecer en otra. En lugar de repetir la información, le dejaremos descubrirlo.

2.1 Herramientas generales

Hay muchas herramientas gratuitas disponibles para tareas múltiples, como la limpieza básica de datos de patentes y la visualización. Aquí destacamos tres herramientas gratuitas.

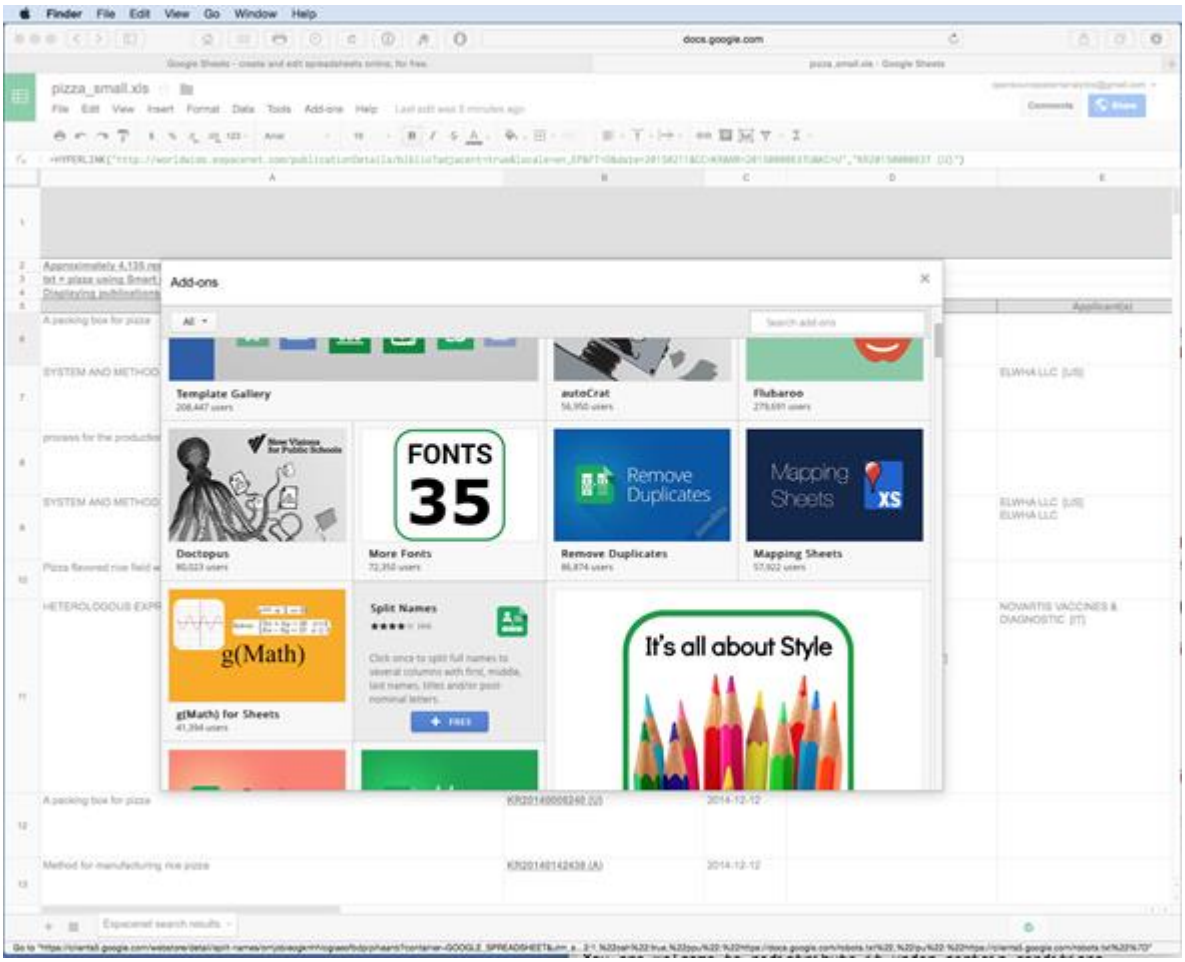
2.1.1 Open Office

Muchos analistas de patentes usarán Excel como un programa predeterminado que incluye la limpieza básica de conjuntos de datos más pequeños. Sin embargo, vale la pena considerar Apache Open Office como una alternativa gratuita. Si bien el análisis de patentes usará la hoja de cálculo (Open Office Calc), también existe una opción de base de datos muy útil como alternativa a Microsoft Access.

- Descargue e instale Apache Open Office para su sistema.
- Consejo: al guardar archivos de hojas de cálculo, elija guardar como **.csv** para evitar situaciones en las que un programa no puede leer los archivos **.odt** predeterminados.

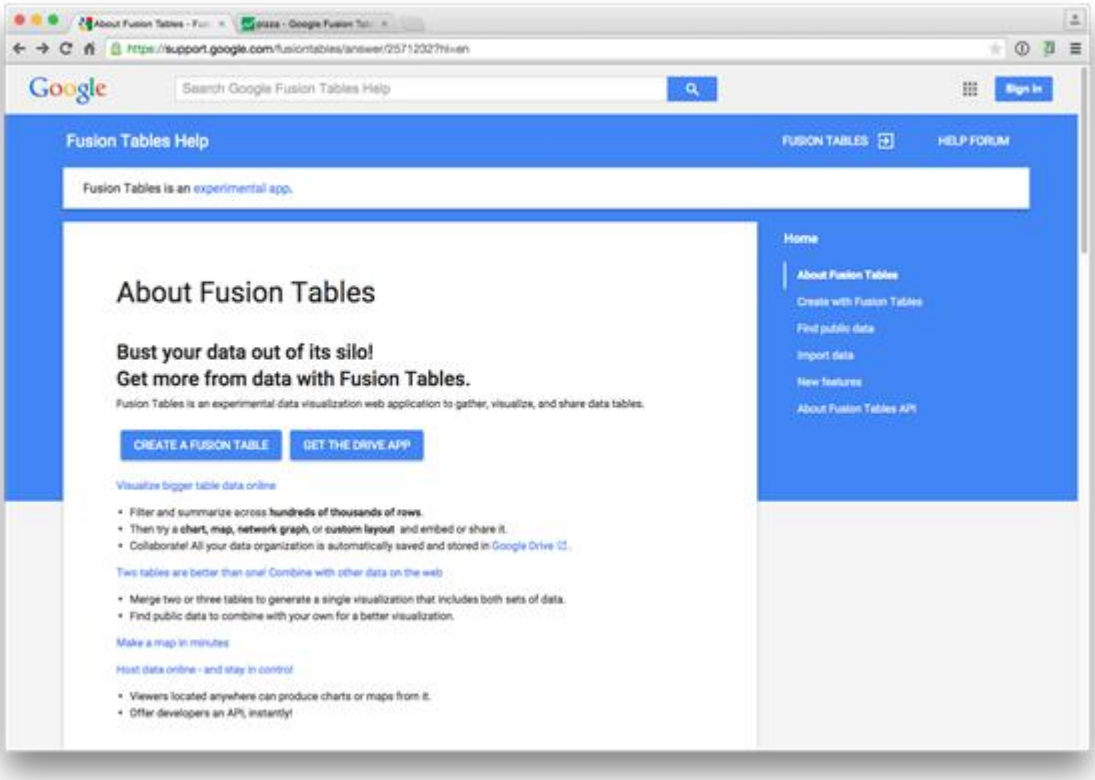
2.1.2 Hojas de Google

Las **Hojas de cálculo de Google** requieren una cuenta gratuita de Google y aquellos que se sienten cómodos con Excel pueden preguntarse por qué vale la pena cambiar. Sin embargo, las hojas de Google se pueden compartir en línea con otras personas y hay una gran cantidad de complementos gratuitos que se pueden usar para ayudar a limpiar datos, como Dividir nombres o Eliminar duplicados, como se muestra a continuación.

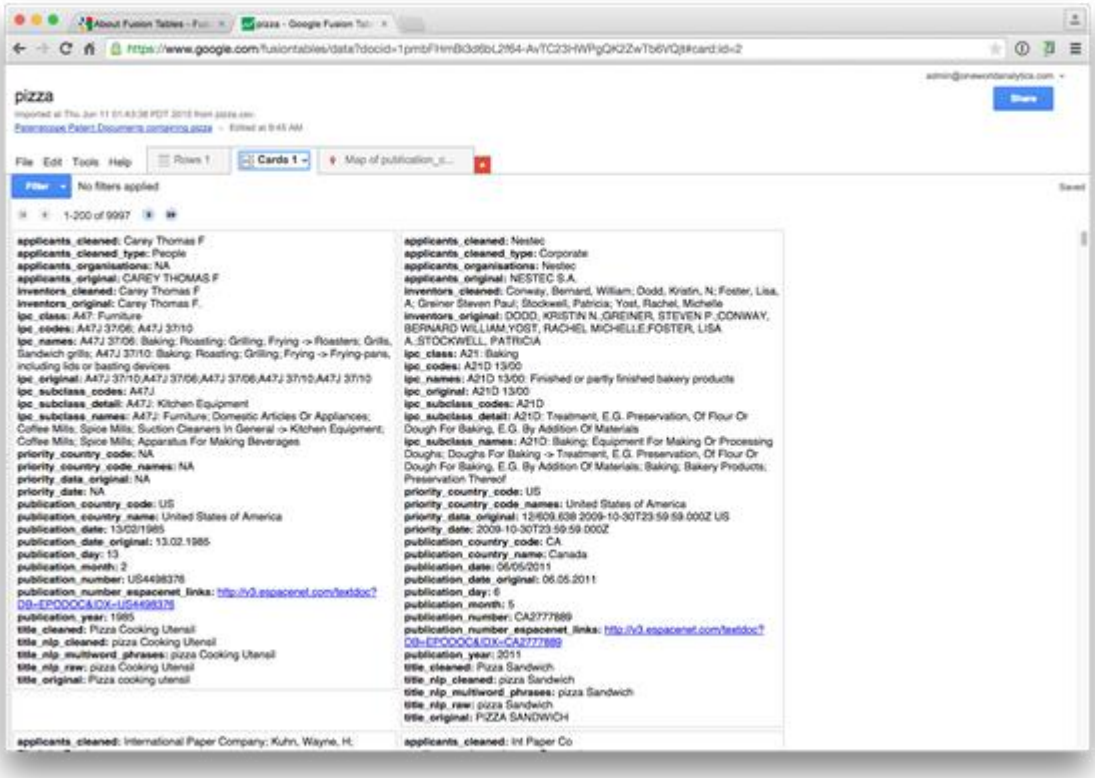


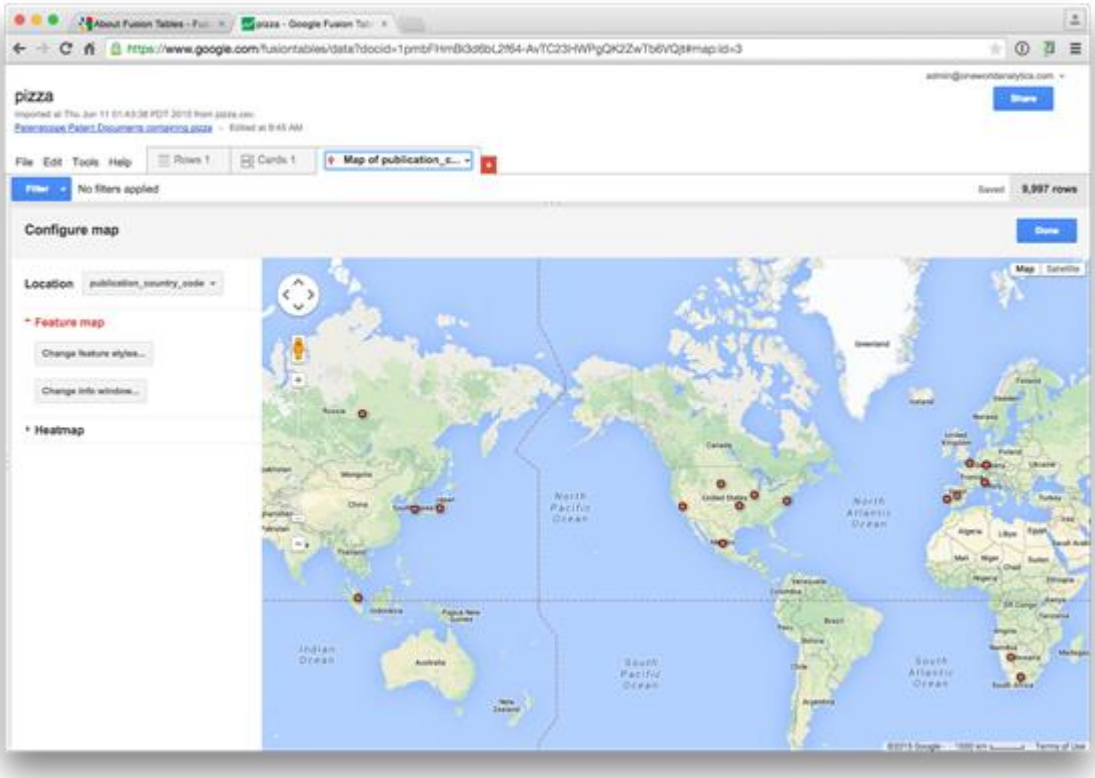
2.1.3 Google Fusion Tables

Las tablas de fusión son similares a las hojas de Google, pero **pueden funcionar con millones de registros**. Sin embargo, vale la pena probar con conjuntos de datos más pequeños para ver si las Tablas de fusión se adaptan a sus necesidades.



Las tablas de fusión se parecen mucho a una hoja de cálculo. Sin embargo, la Tabla también contiene una característica que **permite que cada registro se vea como un todo** y se filtre fácilmente. Puede ser mucho más fácil trabajar con las tarjetas que con el formato de fila estándar donde la información de un registro puede ser difícil de asimilar. **Fusion Tables también intenta usar datos geocodificados para dibujar un mapa de Google** como podemos ver en la segunda imagen a continuación para el País de publicación a partir de un conjunto de datos de patente de muestra.


pizza
File Edit Tools Help
Cards 1
No filters applied
1-200 of 9997
applicants_cleaned: Carey Thomas F
applicants_cleaned_type: People
applicants_organisations: NA
applicants_original: CAREY THOMAS F
inventors_cleaned: Carey Thomas F
inventors_original: Carey Thomas F.
ipc_class: A47: Furniture
ipc_codes: A47J 37/06; A47J 37/10
ipc_names: A47J 37/06: Baking; Roasting; Grilling; Frying -> Roasters; Grills, Sandwich grills; A47J 37/10: Baking; Roasting; Grilling; Frying -> Frying-pans, including lids or basting devices
ipc_original: A47J 37/10;A47J 37/06;A47J 37/06;A47J 37/10;A47J 37/10
ipc_subclass_codes: A47J
ipc_subclass_detail: A47J: Kitchen Equipment
ipc_subclass_names: A47J: Furniture; Domestic Articles Or Appliances; Coffee Mills; Spice Mills; Suction Cleaners In General -> Kitchen Equipment; Coffee Mills; Spice Mills; Apparatus For Making Beverages
priority_country_code: NA
priority_country_code_names: NA
priority_data_original: NA
priority_date: NA
publication_country_code: US
publication_country_name: United States of America
publication_date: 13/02/1985
publication_date_original: 13.02.1985
publication_day: 13
publication_month: 2
publication_number: US4498376
publication_number_espacenet_links: http://v3.espacenet.com/textdoc?DB=EPODOC&IDX=US4498376
publication_year: 1985
title_cleaned: Pizza Cooking Utensil
title_nlp_cleaned: pizza Cooking Utensil
title_nlp_multiword_phrases: pizza Cooking Utensil
title_nlp_raw: pizza Cooking Utensil
title_original: Pizza cooking utensil
applicants_cleaned: Nestec
applicants_cleaned_type: Corporate
applicants_organisations: Nestec
applicants_original: NESTEC S.A.
inventors_cleaned: Conway, Bernard, William; Dodd, Kristin, N; Foster, Lisa, A; Greiner Steven Paul; Stockwell, Patricia; Yost, Rachel, Michelle
inventors_original: DODD, KRISTIN N.;GREINER, STEVEN P.;CONWAY, BERNARD WILLIAM;YOST, RACHEL MICHELLE;FOSTER, LISA A.;STOCKWELL, PATRICIA
ipc_class: A21: Baking
ipc_codes: A21D 13/00
ipc_names: A21D 13/00: Finished or partly finished bakery products
ipc_original: A21D 13/00
ipc_subclass_codes: A21D
ipc_subclass_detail: A21D: Treatment, E.G. Preservation, Of Flour Or Dough For Baking, E.G. By Addition Of Materials
ipc_subclass_names: A21D: Baking; Equipment For Making Or Processing Doughs; Doughs For Baking -> Treatment, E.G. Preservation, Of Flour Or Dough For Baking, E.G. By Addition Of Materials; Baking; Bakery Products; Preservation Thereof
priority_country_code: US
priority_country_code_names: United States of America
priority_data_original: 12/609,639 2009-10-30T23:59:59.000Z US
priority_date: 2009-10-30T23:59:59.000Z
publication_country_code: CA
publication_country_name: Canada
publication_date: 06/05/2011
publication_date_original: 06.05.2011
publication_day: 6
publication_month: 5
publication_number: CA2777689
publication_number_espacenet_links: http://v3.espacenet.com/textdoc?DB=EPODOC&IDX=CA2777689
publication_year: 2011
title_cleaned: Pizza Sandwich
title_nlp_cleaned: pizza Sandwich
title_nlp_multiword_phrases: pizza Sandwich
title_nlp_raw: pizza Sandwich
title_original: PIZZA SANDWICH
applicants_cleaned: International Paper Company; Kuhn, Wayne, H;
applicants_cleaned: Int Paper Co

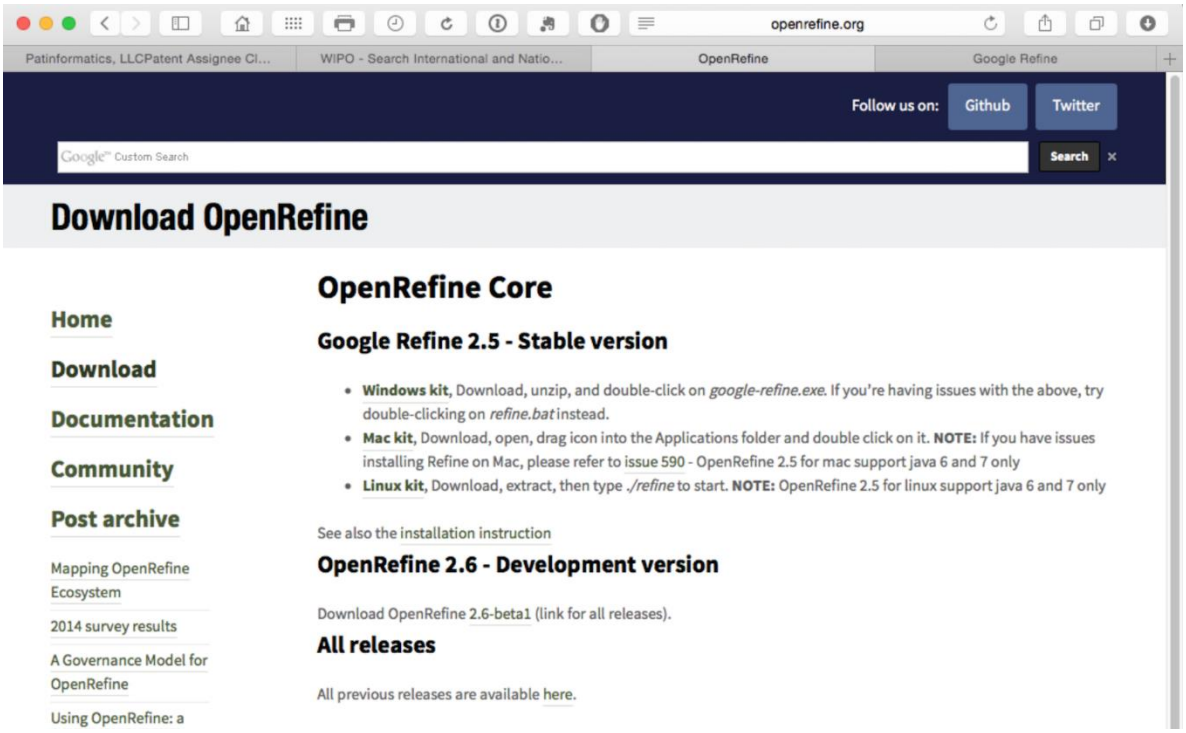

pizza
File Edit Tools Help
Map of publication_c...
No filters applied
9,997 rows
Configure map
Done
Location
publication_country_code
Feature map
Change feature styles...
Change info window...
Heatmap
Map
Satellite
North Pacific Ocean
North Atlantic Ocean
Indian Ocean
South Pacific Ocean
South Atlantic Ocean

2.2 Herramientas Open

2.2.1 Open Refine (anteriormente Google Refine)

Una regla fundamental de análisis de datos y la visualización es: rubbish in = rubbish out. Si sus datos no se han limpiado (corregir o eliminar) en primer lugar, no se sorprenda si los resultados del análisis o la visualización son basura.

Un capítulo detallado está disponible aquí sobre el uso de Open Refine , anteriormente Google Refine, para limpiar datos de patentes. Para el análisis de patentes, Open Refine es una importante herramienta gratuita para limpiar nombres de solicitantes e inventores.



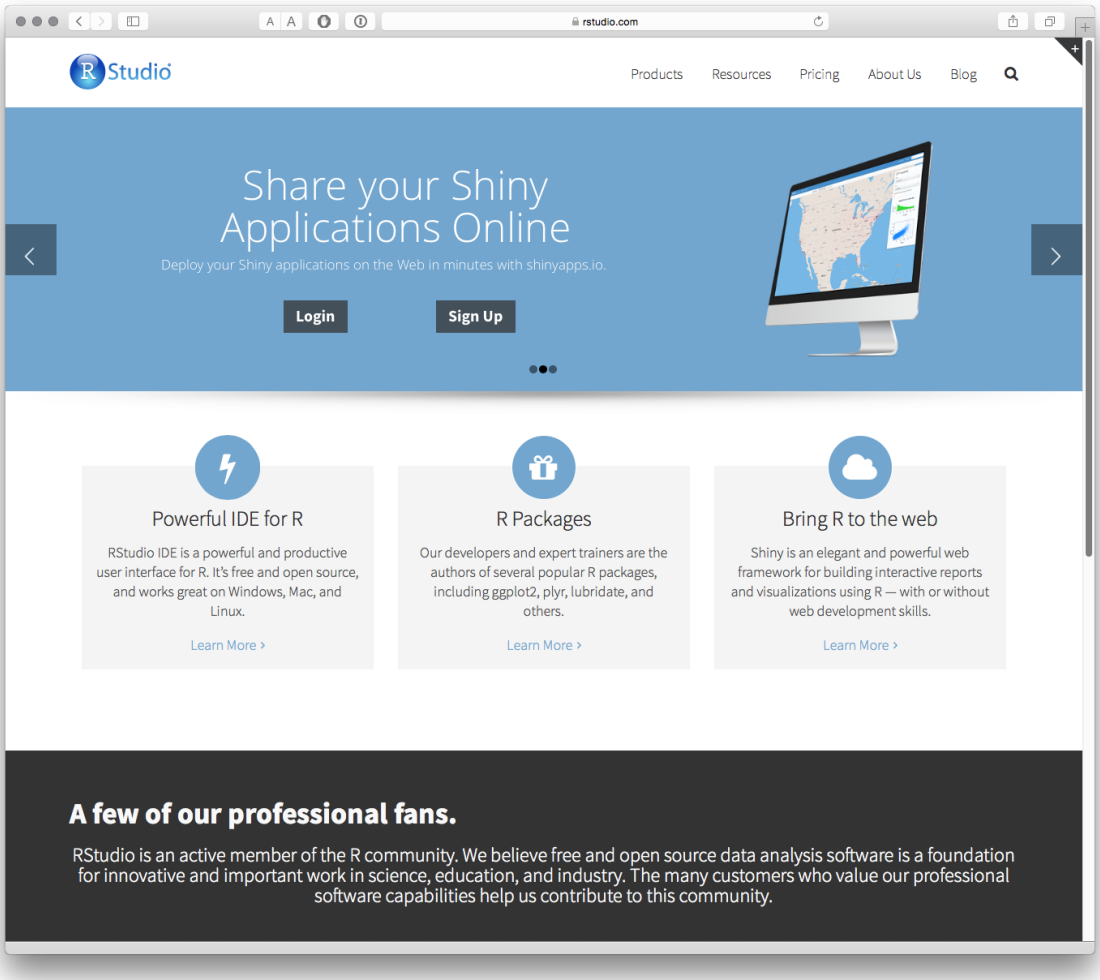
Una serie de plataformas proporcionan facilidades de limpieza de datos y es posible realizar bastantes limpiezas básicas en Open Office o Excel. Open Refine es la herramienta más accesible para la limpieza oportuna de los campos de nombre de patente. En particular, es muy útil para dividir y limpiar miles de nombres de solicitantes de patentes e inventores.

2.3 Minería de datos

Hay un número cada vez mayor de herramientas de minería de datos por ahí. Aquí hay algunos de los que han llamado nuestra atención con las herramientas adicionales que se enumeran a continuación.

2.3.1 RStudio

Una herramienta muy poderosa para trabajar con datos y visualizar datos usando R y luego escribir sobre ellos (este capítulo, y el Manual más amplio, están escritos completamente en Rmarkdown con RStudio). Si bien la curva de aprendizaje con R puede ser intimidante, se debe hacer un gran esfuerzo para hacer que R sea accesible a través de tutoriales como los de DataCamp , webinars , R-Bloggers y Stack Overflow y cursos universitarios gratuitos como el conocido programa de programación de John Hopkins University R Curso de Coursera . De hecho, al igual que con Python, hay tanto soporte para usuarios en diferentes niveles que es difícil sentirse solo cuando se usa R y RStudio.



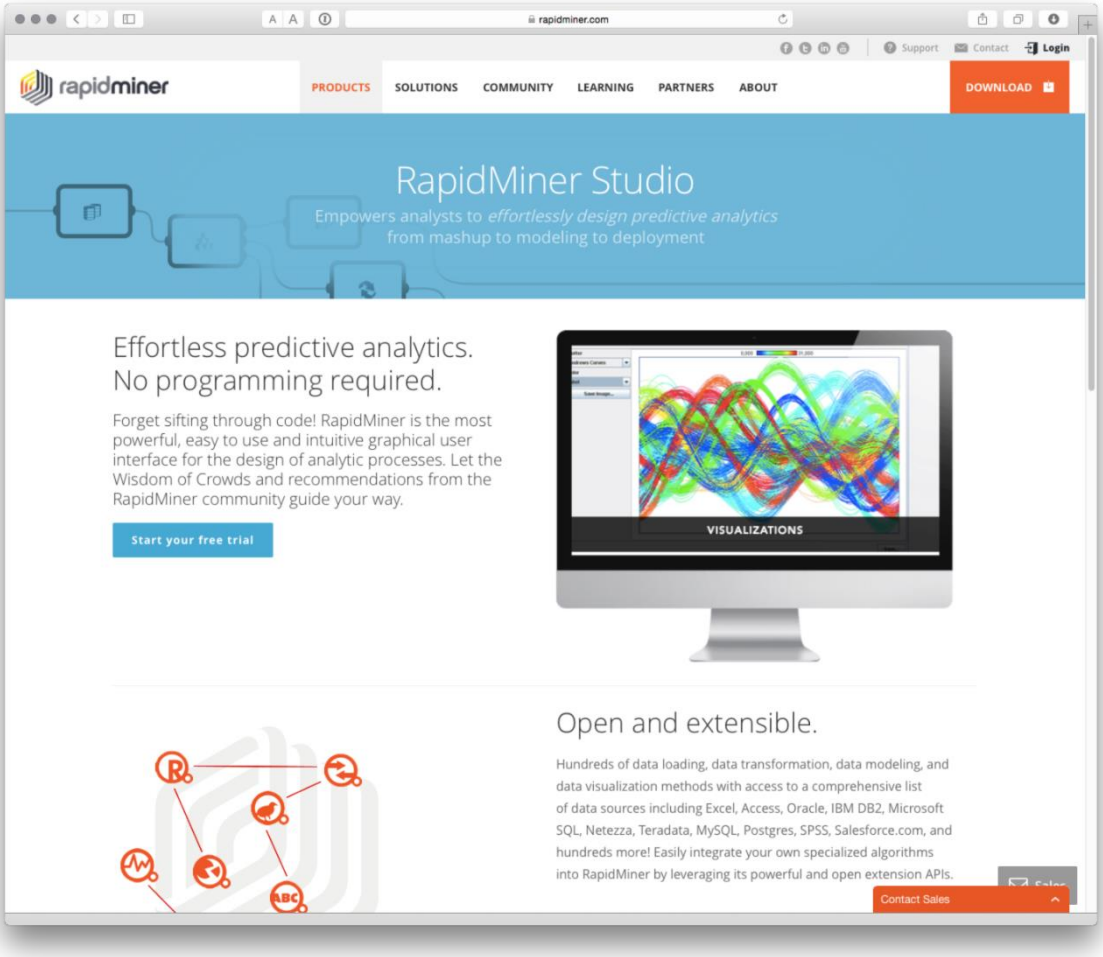
Para comenzar con R, descargue RStudio para su plataforma siguiendo estas instrucciones y asegurándose de instalar R desde el enlace proporcionado.

Si eres completamente nuevo en R, entonces DataCamp es un buen lugar para comenzar. El curso gratuito de programación R de la Universidad John Hopkins en Coursera también es muy bueno. El curso de la Universidad John Hopkins está acompañado por el paquete tutorial Swirl que se puede instalar usando `install.packages (" swirl ") cuando instaló R. Este es un activo real al comenzar.

Al desarrollar este Manual, nos enfocamos principalmente en desarrollar recursos con R. Sin embargo, enfatizamos que Python también puede ser importante para sus necesidades. Para una discusión reciente sobre las fortalezas y debilidades de R y Python, vea este artículo de Datacamp sobre la Data Science Warsexcelente infografía que lo acompaña .

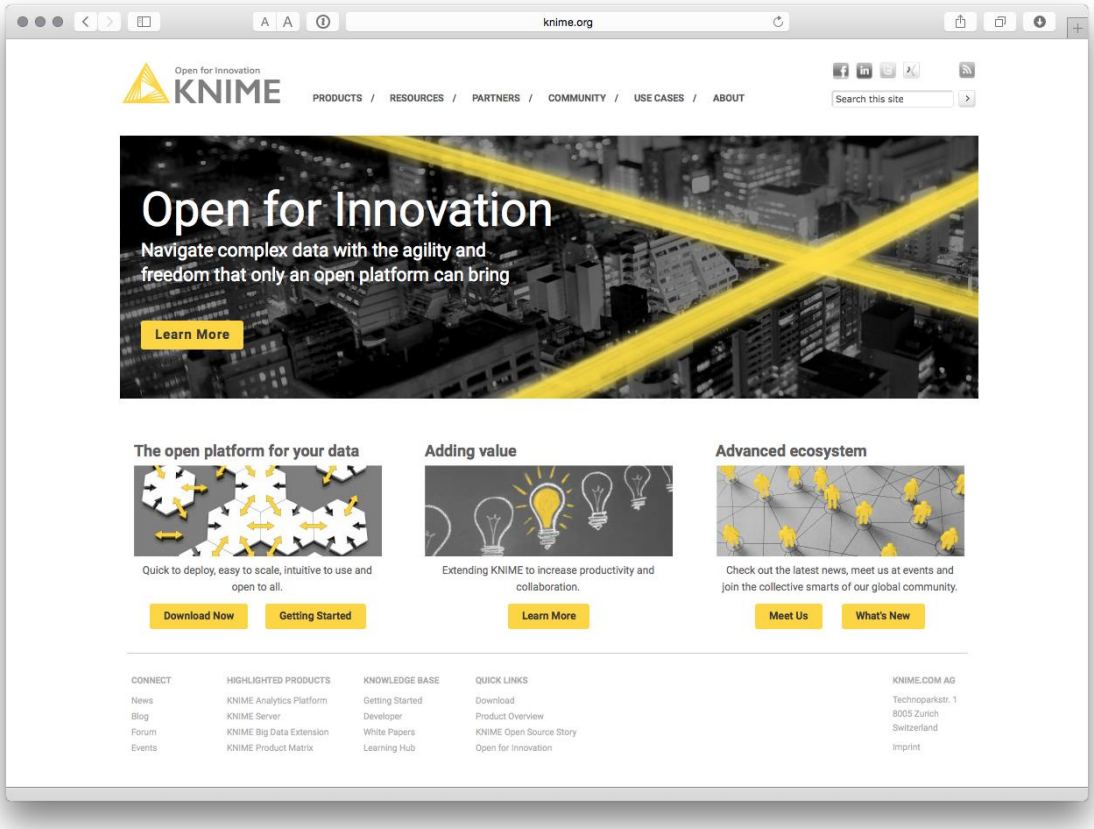
2.3.2 RapidMiner Studio

Viene con un servicio gratuito y una variedad de planes pagados por niveles. RapidMiner se centra en el aprendizaje automático, la minería de datos, la minería de textos y el análisis.



2.3.3 el [Knime]

Una plataforma abierta para la minería de datos.



A continuación se describen otras herramientas de extracción de datos (como WEKA y NLTK en Python). Si desea explorar otro software de minería de datos, pruebe este artículo para obtener algunas ideas.

2.4 Visualización de datos

Si eres nuevo en la visualización de datos, te sugerimos que te interese el trabajo de **Edward Tufte en la Universidad de Yale y** su famoso libro La visualización visual de información cuantitativa . Su crítica de los usos y abusos de Powerpoint también es entretenida y perspicaz. El trabajo de Stephen Few, como Show Me the Numbers: Designing Tables and Graps to Enlighten, también es popular.

Recuerde que la visualización de datos es, ante todo, la comunicación con un público. Eso implica elecciones sobre cómo comunicarse y encontrar maneras de comunicarse claramente. En muchos casos, el resultado del análisis y visualización de patentes será un informe y una presentación. La crítica de Tufte de las presentaciones en powerpoint debería ser obligatoria para los presentadores. También le recomendamos que eche un vistazo al Resonate de Nancy Duarte para obtener ideas sobre cómo pulir las presentaciones y contar historias. Es posible que

el estilo no sea adecuado para todos, pero Resonate contiene mensajes e ideas muy útiles. En un entorno sin conexión, considere el Atlas de la Ciencia de Katy Borner : Visualizar lo que sabemoscomo una excelente guía para la historia de las visualizaciones de la actividad científica, incluidas las visualizaciones pioneras de la actividad de patentes. Tenga en cuenta que la visualización efectiva requiere práctica y es un camino bastante transitado.

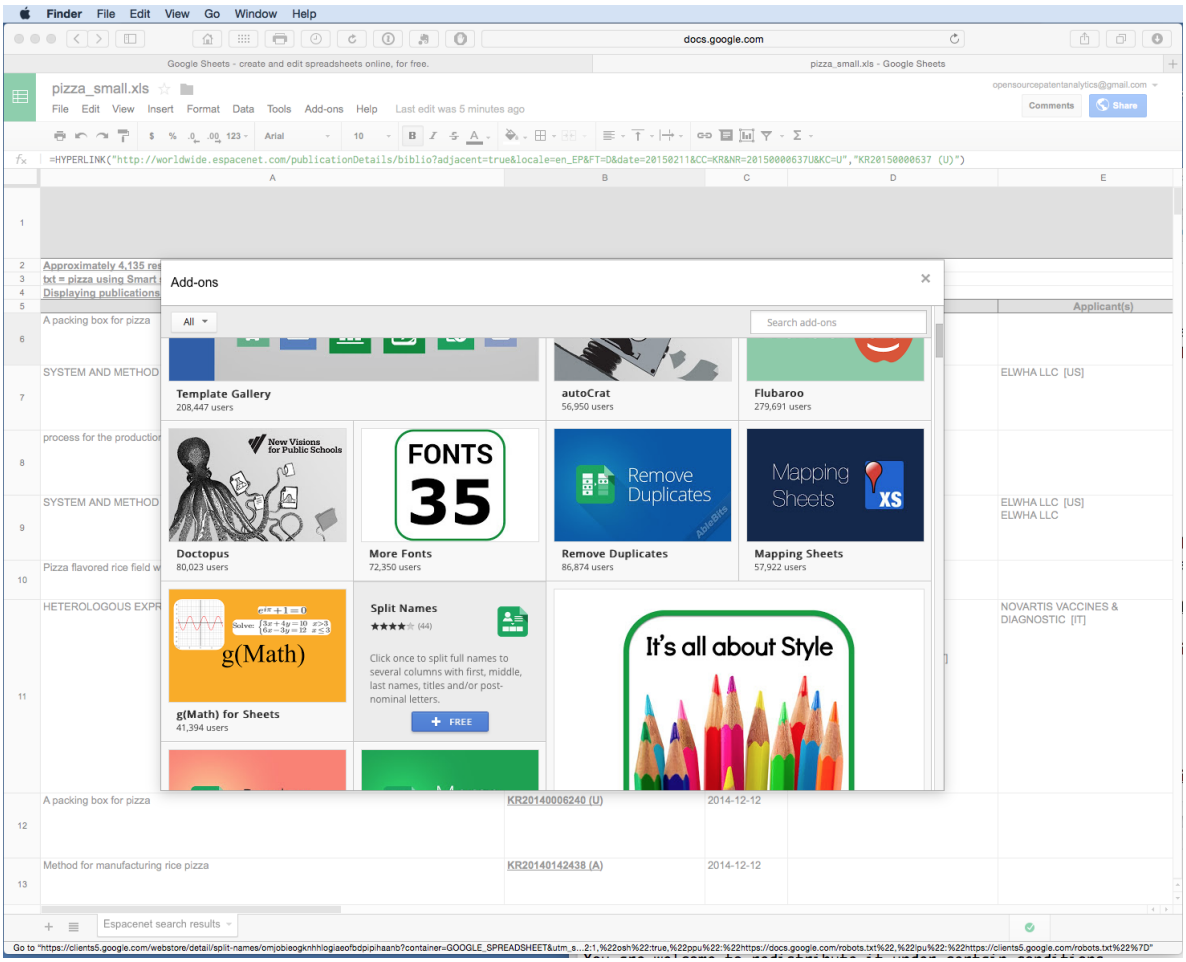
Hay muchas opciones para las herramientas de visualización de datos y la cantidad de herramientas está creciendo rápidamente. Para análisis de negocios, Gartner proporciona un útil (pero basado en suscripción) Magic Quadrant para Business Intelligence and Analytics que busca identificar a los líderes en el campo. Estos tipos de informes pueden ser útiles para localizar y acercarse a las compañías y verificar si existe una versión gratuita del software (que no sea una breve prueba gratuita).

Sugerimos pensar cuidadosamente sobre sus necesidades y la curva de aprendizaje involucrada. Por ejemplo, si tiene conocimientos limitados de programación (o no tiene tiempo o ganas de aprender) elija una herramienta que haga el trabajo en gran medida por usted. Si ya tiene experiencia con javascript, Java, R o Python, o similar, elija una herramienta con la que se sienta más cómodo. En particular, preste atención a las herramientas con una API (interfaz de programación de aplicaciones) en una variedad de idiomas (como Python o R) que puedan satisfacer sus necesidades.

Si usted es completamente nuevo en la visualización de datos, Tableau Public y nuestro capítulo explicativo son un buen lugar para aprender sin saber nada sobre programación. Algunas otras herramientas en esta lista son similares a Tableau Public (en parte porque Tableau es el líder del mercado). También le proporcionaremos algunos consejos para los sitios de visión general de la visualización al final de esta sección, donde podrá encontrar información sobre lo que es nuevo e interesante en la visualización de datos.

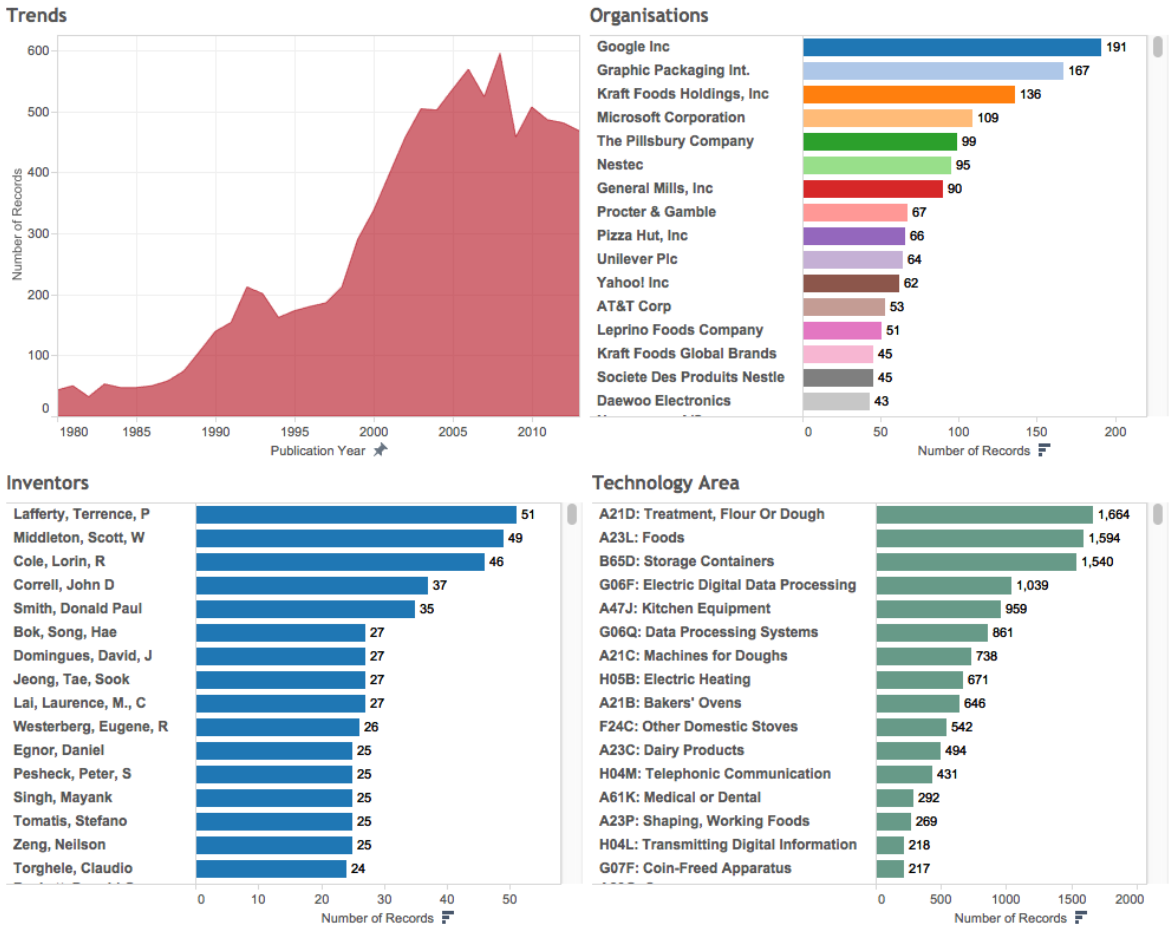
2.4.1 Google Charts

- Cree una cuenta de Google para acceder a las hojas de cálculo de Google y otros programas de Google
- Echa un vistazo a la galería de gráficos de Google y la API
- Para obtener una descripción general del uso de Google Charts en R, consulte el GoogleVispaquete y sus ejemplos aquí
- Para obtener información general sobre el uso de Google Charts con Python, consulte google-chartwrapper o Python Google Charts



2.4.2 [Tableau Public]

Un capítulo detallado sobre cómo comenzar con el análisis y visualización de patentes utilizando Tableau Public está disponible [aquí] . Cuando se han limpiado los datos de patentes, Tableau Public es una forma eficaz de desarrollar cuadros de mando interactivos y mapas con sus datos y combinarlos con otras fuentes de datos. Tenga en cuenta que los datos públicos de Tableau son, por definición, públicos y no deben utilizarse con datos confidenciales.



El libro se puede ver en línea aquí .

2.4.3 R y RStudio

R es un lenguaje de programación estadística para trabajar con todo tipo de diferentes tipos de datos. También cuenta con potentes herramientas de visualización que incluyen packag es una interfaz con Google Charts, Plotly y otros. Si está interesado en utilizar R, le sugerimos que utilice RStudio, que puede descargar aquí . El Manual completo de análisis de patentes de código abierto de la OMPI se escribió en RStudio utilizando Rmarkdown para imprimir los artículos para la web, .pdf y presentaciones. Como esto sugiere, no se trata simplemente de la visualización de datos. Para comenzar con R y RStudio, pruebe los tutoriales gratuitos en DataCamp . Cubriremos R con más detalle en otros capítulos y artículos en línea.

Como parte de un enfoque descrito como The Grammar of Graphicsinspirado por el trabajo de Leland Wilkinson , los desarrolladores de RStudio y otros han creado paquetes que proporcionan formas muy útiles de visualizar y mapear datos. Los

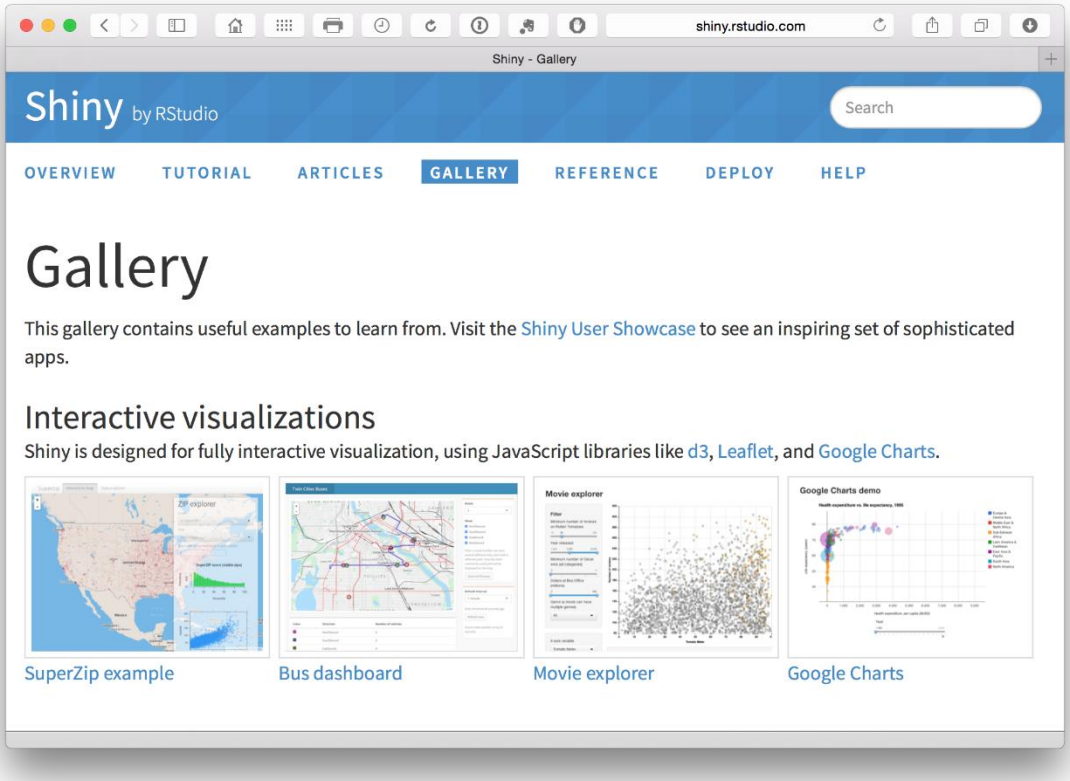
enlaces a continuación lo llevarán a la documentación de algunos de los paquetes de visualización de datos más populares.

1. ggplot2
2. ggvis
3. ggmap
4. googlevis

Cubriremos ggplot2y ggvis con mayor profundidad en futuros capítulos. Hasta entonces, para comenzar, vea los capítulos ggplot2 sobre R-Bloggers y aquí para ggvis . Datacamp ofrece un tutorial gratuito sobre el uso de ggvis que se puede acceder aquí . Para obtener una descripción más amplia de algunos de los mejores paquetes de R, vea la impresionante lista de R de Qin Wenfeng.

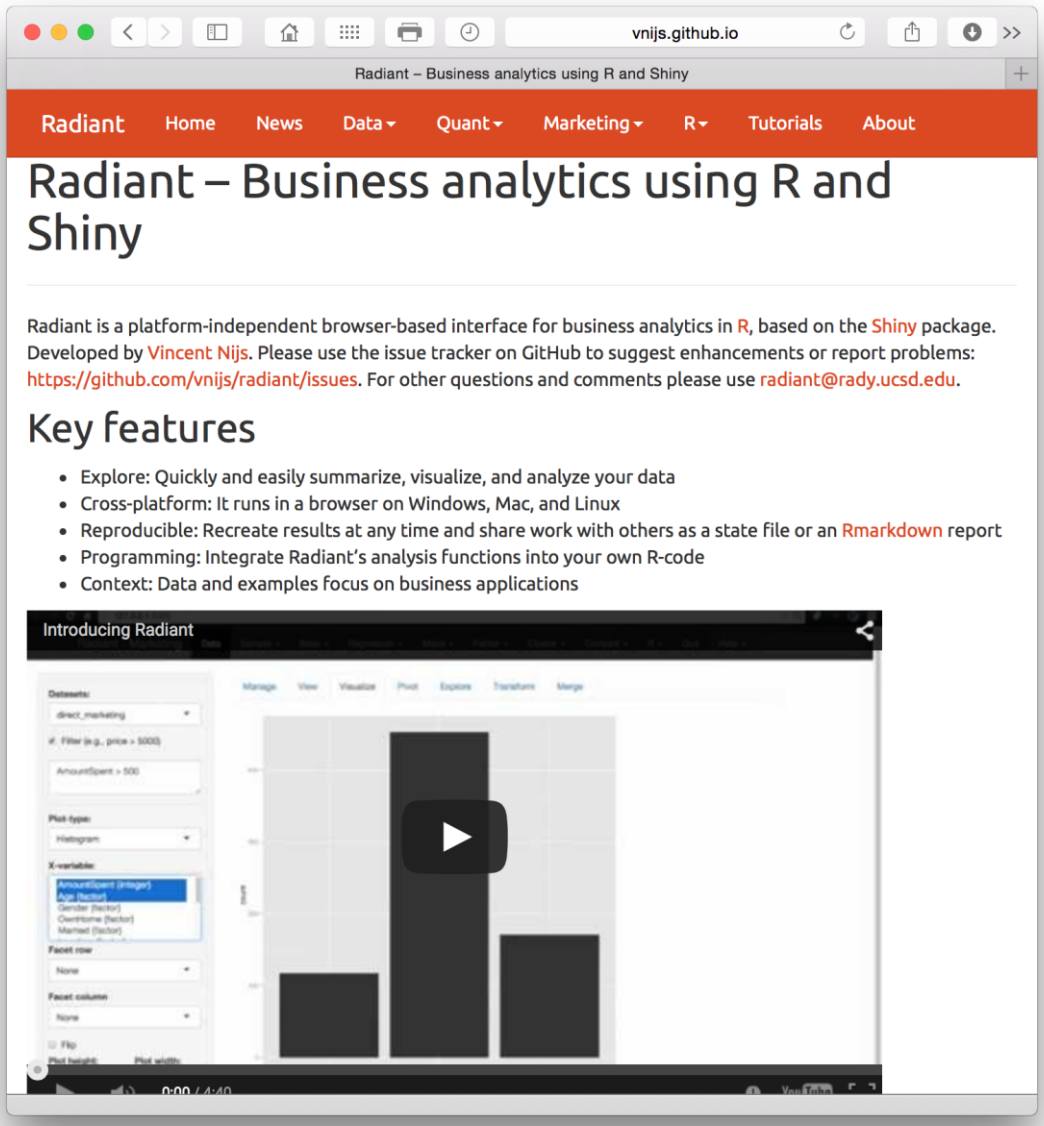
2.4.4 Brillante

Shiny from RStudio es un marco de aplicación web para R. Lo que significa es que puede generar tablas y datos visuales de R, como los de las herramientas mencionadas anteriormente a la web.



Las aplicaciones Shiny para usuarios de R permiten la creación de aplicaciones interactivas en línea (hasta 5 gratis). Vea la Galería para ejemplos. Ver RBloggers para más ejemplos y tutoriales.

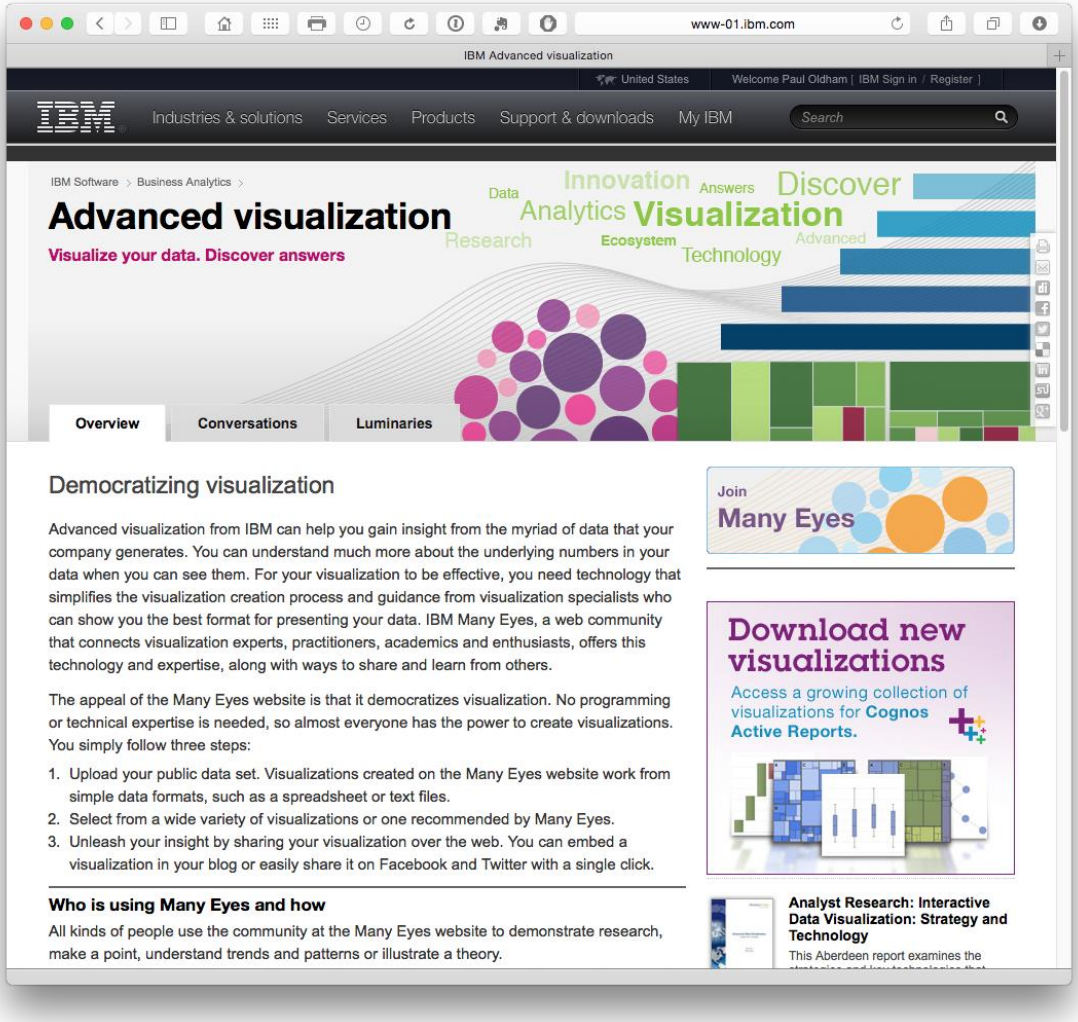
Radiant es una plataforma basada en navegador para análisis de negocios en R. Se basa en Shiny (arriba) pero está específicamente enfocada en los negocios.



Para una serie de videos de inicio en Radiant vea aquí .

2.4.5 IBM Many Eyes

Debe registrarse para obtener una cuenta gratuita para entender realmente de qué se trata, pruebe esta página y seleccione registrarse en la parte superior derecha.



2.4.6 Otras herramientas de visualización

- Tulip : marco de visualización de datos en C ++
- SigmaJS : biblioteca de JavaScript dedicada al dibujo gráfico. Permite la creación de gráficos interactivos estáticos y dinámicos.
- Kendo UI : Crea widgets para visualizaciones sensibles.
- Línea de tiempo : Un KnightLab (universidad del noroeste) es una herramienta que permite la creación de líneas de tiempo interactivas y está disponible en 40 idiomas.

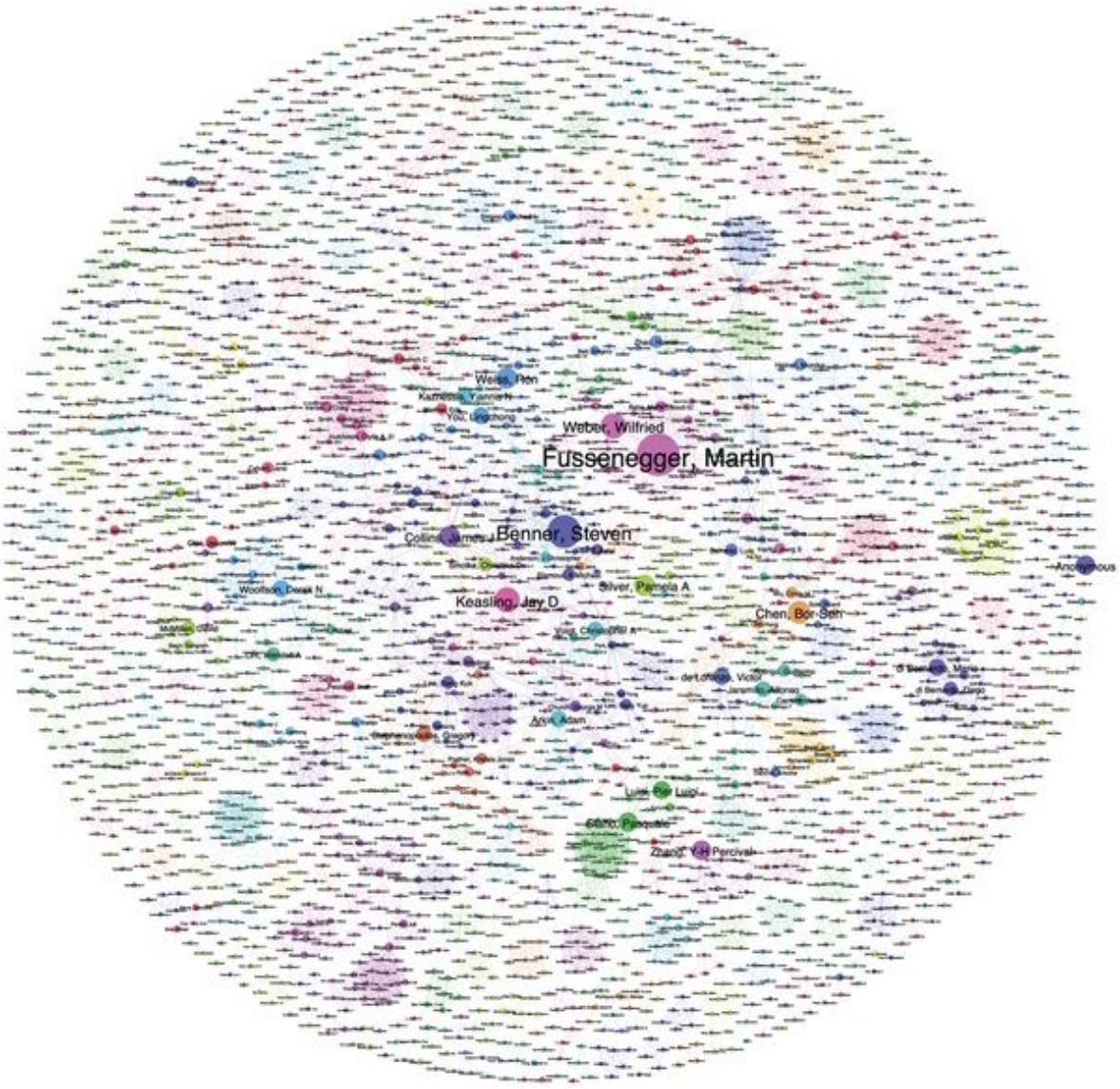
- Proyecto Miso : kit de herramientas de código abierto que facilita la creación de narración interactiva y visualización de datos
- Sci2 : Un conjunto de herramientas para el estudio de la ciencia.
- Widgets de Simile Widgets web para contar historias como un resultado del Proyecto SIMILE en el MIT.
- jqPlot . Un código abierto basado en jQuery Plotting Plugin.
- Dificultad para Timelines (servicios gratuitos y premium).

Para obtener herramientas e ideas de visualización adicionales, visite visualizing.org y Open Data Tools .

2.5 Visualización de la red

El software de visualización de redes es una herramienta importante para visualizar actores en un campo de la ciencia y la tecnología y, en particular, las relaciones entre ellos. Para el análisis de patentes, se puede utilizar para una variedad de propósitos que incluyen:

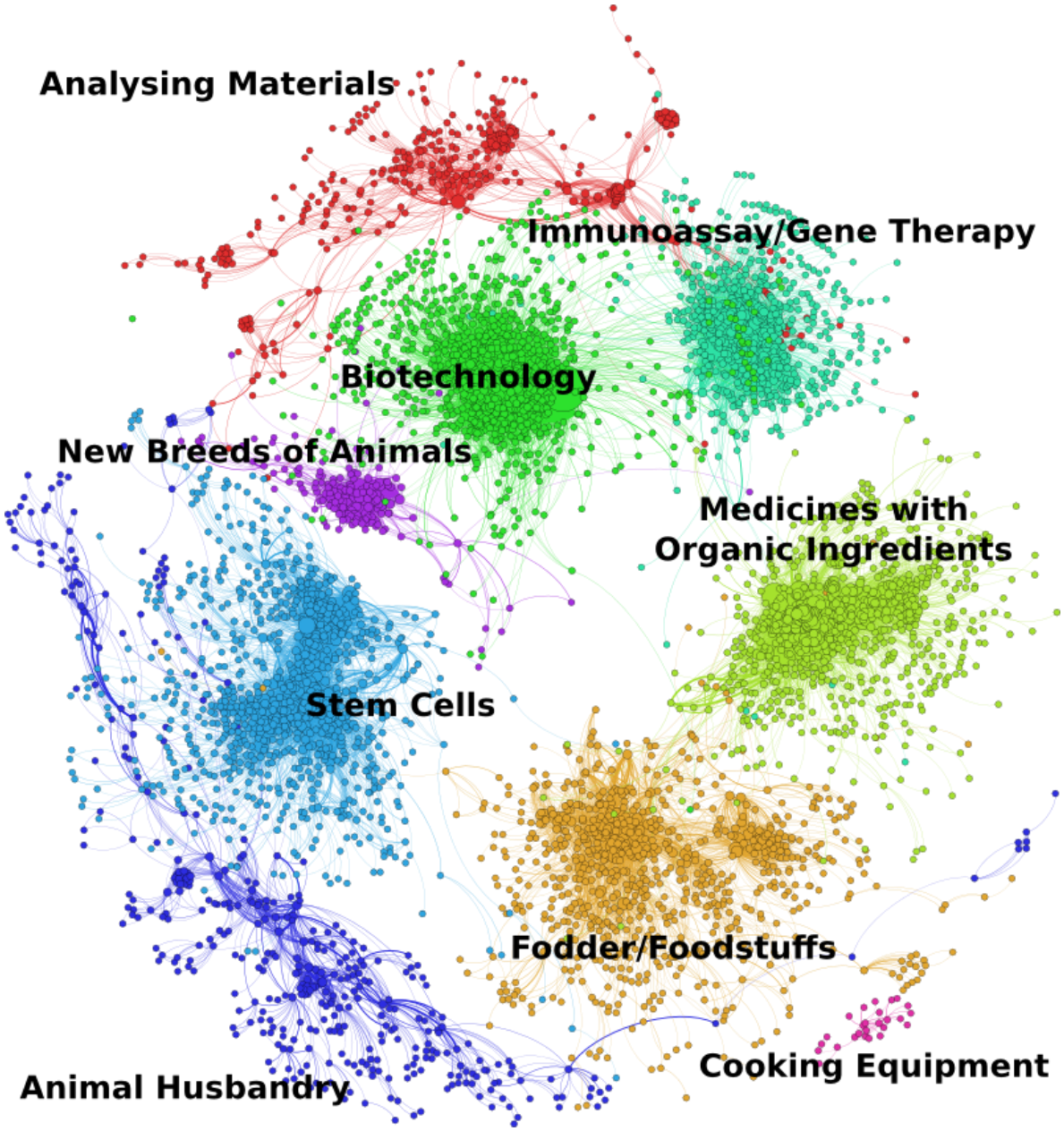
1. Visualización de redes de solicitantes e inventores en un campo particular o investigadores científicos. Un ejemplo de este tipo de trabajo para biología sintética se encuentra aquí para una red de aproximadamente 2,000 autores de artículos sobre biología sintética.



2. Visualización de áreas de tecnología y sus relaciones utilizando la Clasificación Internacional de Patentes y la Clasificación Cooperativa de Patentes (CPC). El trabajo anterior en la OMPI fue pionero en el uso del análisis de redes de patentes a gran escala para identificar el panorama de patentes para los recursos zoogenéticos .

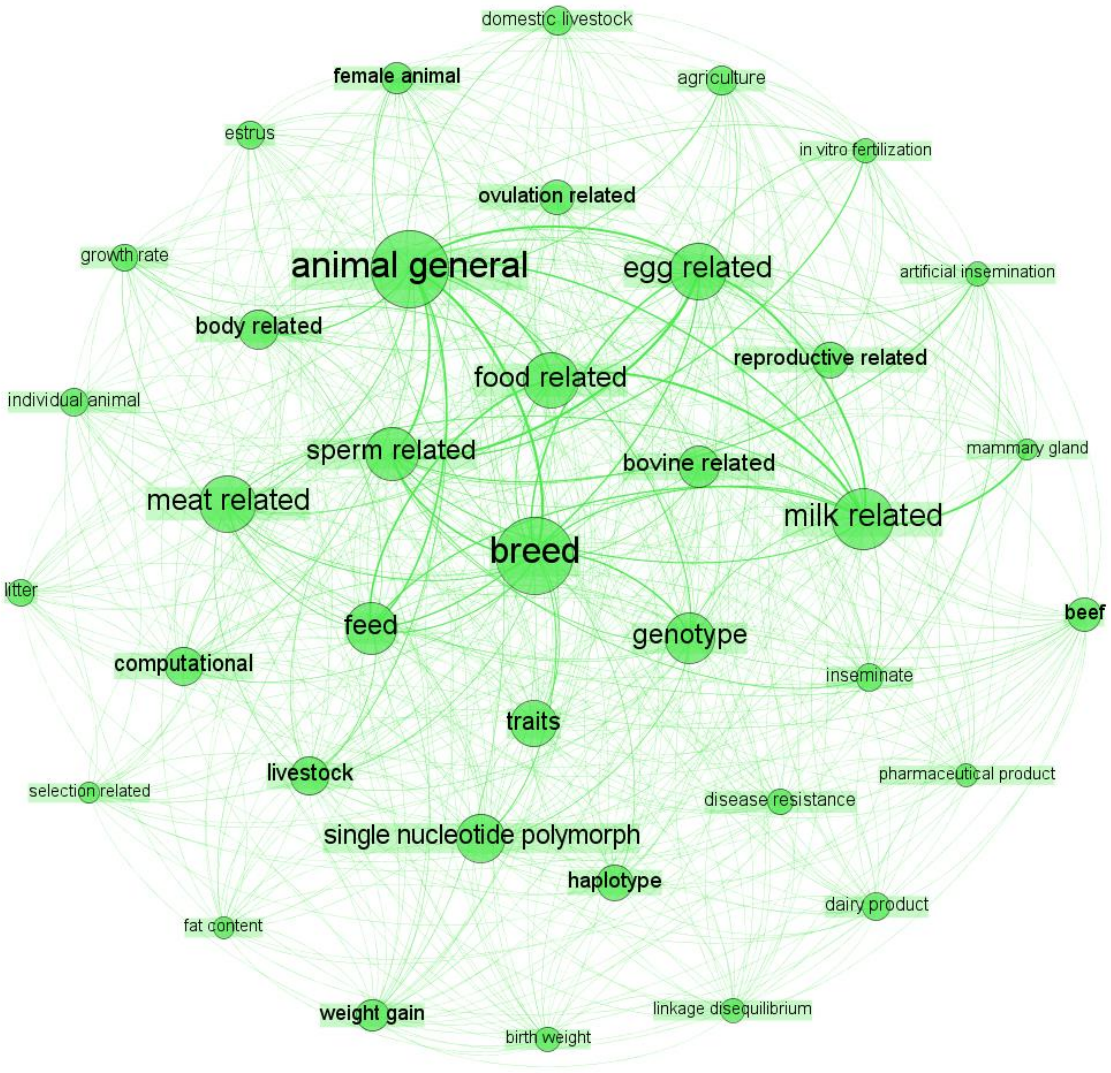
La imagen a continuación muestra un mapa de la red de códigos de clasificación de patentes cooperativas y códigos de clasificación de patentes internacionales para decenas de miles de documentos de patentes que contienen referencias a una variedad de animales de granja (vacas, cerdos, ovejas, etc.). Los puntos son códigos de CPC / IPC que describen áreas de tecnología. Los clústeres muestran documentos estrechamente vinculados que comparten los mismos códigos que luego pueden describirse como 'módulos' o clústeres. Los autores del informe del

paisaje sobre recursos genéticos animales utilizaron esta red como una herramienta exploratoria para extraer y examinar los documentos en el grupo de relevancia. Se desecharon los grupos distantes, como el equipo de cocina y la cría de animales (alojamiento de animales, etc.). Más tarde, los autores utilizaron el mapeo de redes para explorar y clasificar los grupos individuales.



3. Visualizar redes de términos clave en documentos de patente y sus relaciones con otros términos como parte de la exploración y el refinamiento del análisis. En este caso, los autores han agrupado términos similares entre sí utilizando palabras derivadas para comprender el contenido de las nuevas

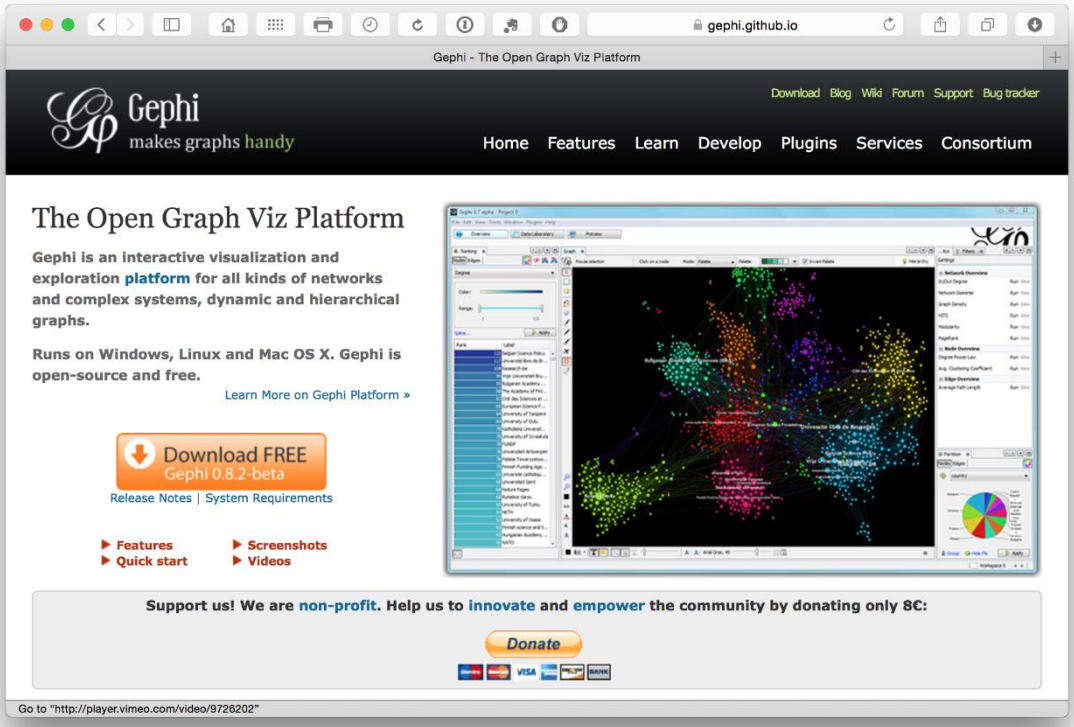
razas de animales agrupados anteriormente en relación con la cría de animales.



Como tal, la visualización de la red puede verse como una herramienta exploratoria para definir el objeto de interés y como el resultado final (por ejemplo, una red definida de actores en un área específica).

2.5.1 Gephi

Gephi es un software de generación de red de código abierto basado en Java. Puede hacer frente a grandes conjuntos de datos (en función de su computadora) para producir visualizaciones potentes.

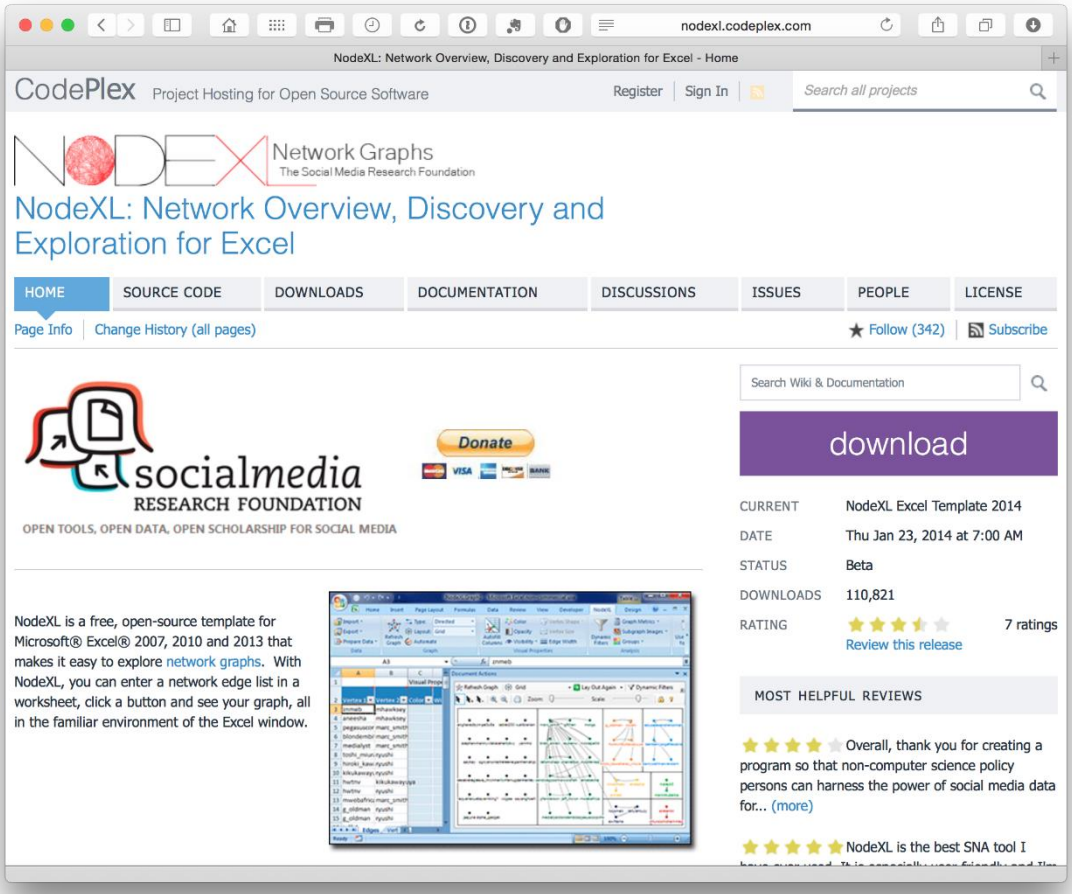


Un problema que puede surgir, en particular por los usuarios de Mac, son los problemas para instalar la versión Java correcta. Este problema parece haberse resuelto con la última versión, versión 0.9.

Para crear .gexfarchivos de red en R, pruebe el paquete gexf y el código de ejemplo y el código fuente aquí . En Python, pruebe la biblioteca pygexf y para cualquier otra cosa como Java, Javascript C ++ y Perl, visite gexf.net .

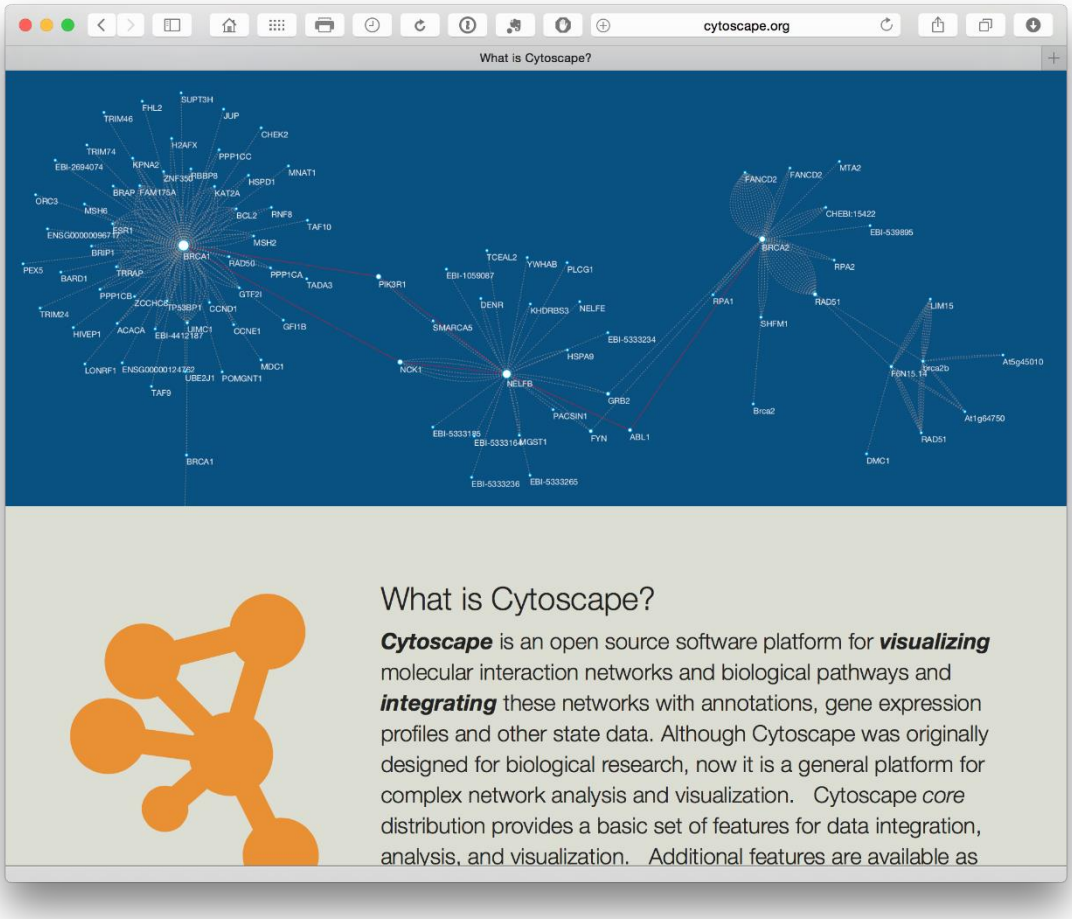
2.5.2 NodeXL

Para los usuarios de Excel, NodeXL es un complemento que se puede usar para visualizar redes. Funciona bien.



2.5.3 Cytoscape

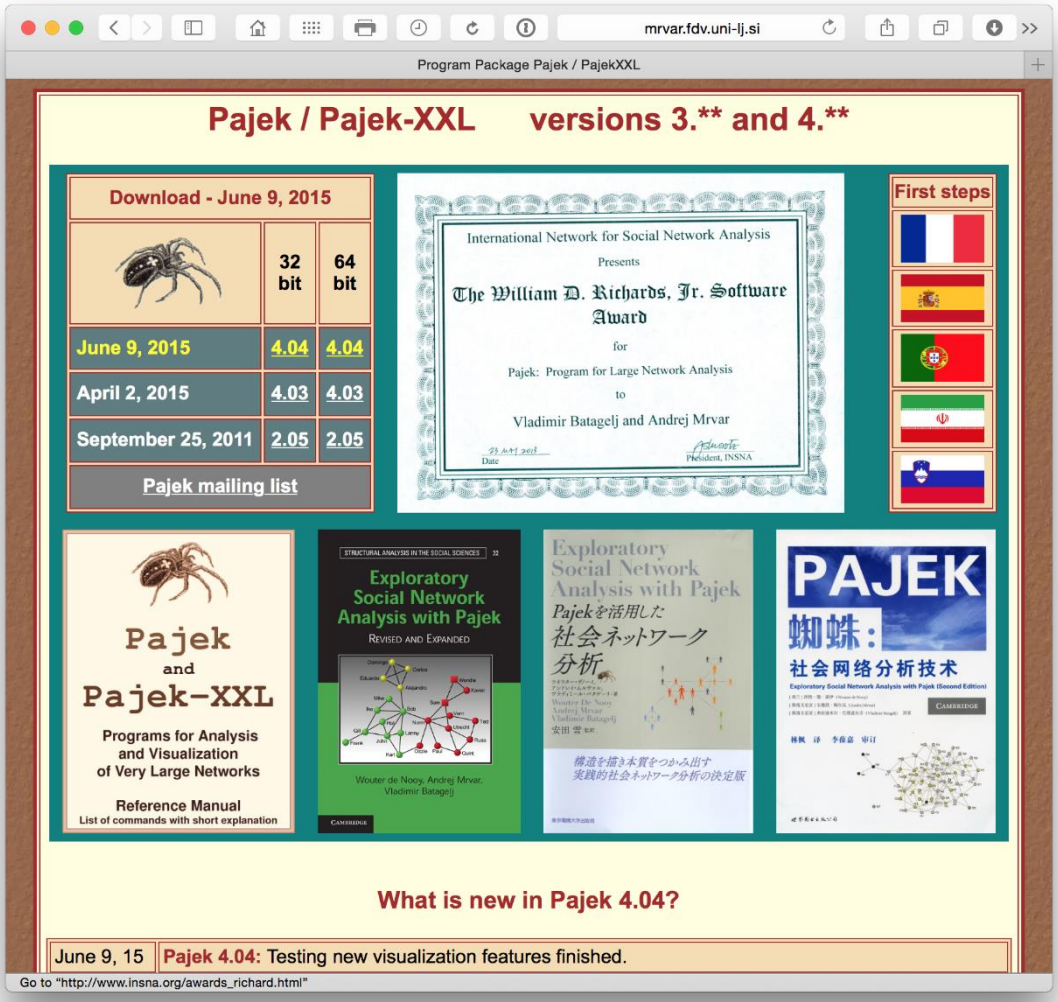
Cytoscape es otro programa de visualización de red. Originalmente fue diseñado para la visualización de redes biológicas e interacciones pero, al igual que con muchas otras herramientas de bioinformática, se puede aplicar a una amplia gama de tareas de visualización.



Principalmente tenemos experiencia con el uso de Gephi (arriba) pero vale la pena explorar Cytoscape. Cytoscape funciona con Windows, Mac y Linux.

2.5.4 Pajek

Esta es una de las herramientas de red gratuitas más antiguas y establecidas, y es solo para Windows (o se ejecuta a través de una Máquina Virtual). Es ampliamente utilizado en bibliometría y puede manejar grandes conjuntos de datos. Es una cuestión de preferencia personal, pero herramientas como Gephi pueden estar reemplazando a Pajek porque son más flexibles. Sin embargo, es posible que Pajek tenga una ventaja en cuanto a precisión, facilidad de reproducibilidad y la importante capacidad para guardar fácilmente el trabajo que Gephi puede carecer como programa Beta.



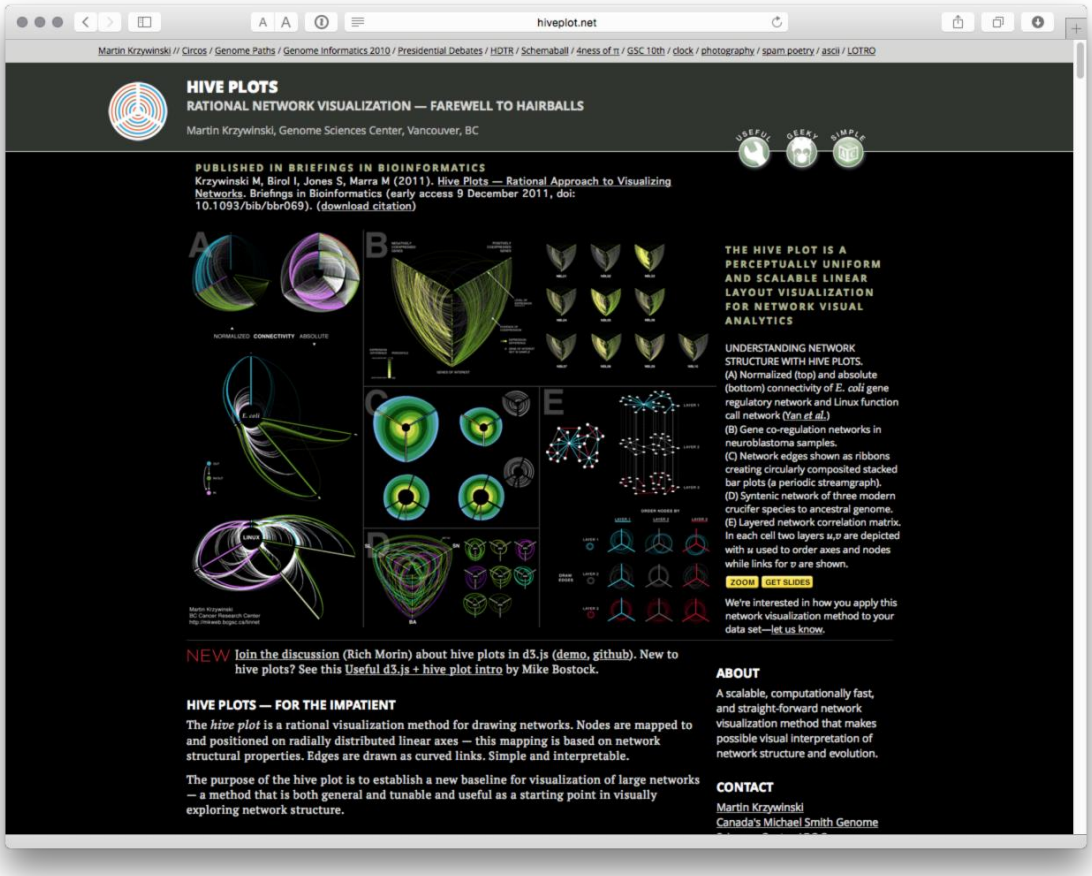
Los datos también se pueden exportar de Pajek a Gephi para aquellos que prefieren la apariencia de Gephi.

2.5.5 Visor VOS

VOS Viewer de la Universidad de Leiden es similar a Gephi y Cytoscape, pero también presenta diferentes tipos de paisaje (a diferencia de los nodos de red puros y las imágenes de borde). La última versión también puede hablar con Gephi y Cytoscape. **Vale la pena probar diferentes opciones de visualización visual y su capacidad para manejar datos bibliográficos de Web of Science y Scopus.**

2.5.6 Hive Plots

No estamos completamente seguros de qué hacer con Hive Plots. Sin embargo, tenemos mucha simpatía con los objetivos. El objetivo de la visualización en red debe ser aclarar el complejo ... no "wow, mira, hice algo que se parece a un espagueti" (aunque eso normalmente es parte del proceso). Por lo tanto, encontramos Hive Plots desarrollados por Martin Krzywinski en el Genome Sciences Center de la Agencia de Cáncer de BC.



Diseñado para redes grandes, hay paquetes para diagramas de Hive en Python a través de pyveplot y hiveplot . Para R hay HiveR con documentación disponible en CRAN aquí .

Al cerrar este análisis de las herramientas de mapeo de red, también es importante tener en cuenta que las visualizaciones de red deben exportarse como imágenes. Esto significa que hay requisitos adicionales para el software de manejo de imágenes. Las herramientas de código abierto como el Programa de manipulación de imágenes GNU o GIMP son perfectamente adecuadas y fáciles de usar para el manejo de imágenes. Cuando se utilicen etiquetas, se debe prestar especial atención a delinear el texto para garantizar la coherencia de la visualización en diferentes equipos. Este tipo de tareas se pueden realizar en herramientas como GIMP.

Para otras fuentes de visualización de red vea FlowingData . Prueba también la complejidad visual y la visualización de datos para obtener fuentes de inspiración.

2.6 Infografías

Las infografías son cada vez más parte del conjunto de herramientas de comunicación. Son particularmente útiles para comunicar los resultados de la investigación en una forma informativa fácilmente digerible. El Proyecto de paisaje de patentes de la OMPI ha desarrollado una serie de infografías con lo último en el Informe de paisaje de patentes de recursos genéticos animales y Dispositivos y tecnologías de asistencia .

La creciente popularidad de la infografía ha sido testigo del aumento de una gama de servicios en línea que incluyen servicios gratuitos. En la mayoría de los casos, estos tendrán limitaciones, como la cantidad de iconos, etc., que se pueden usar en un gráfico. Sin embargo, como un sector en crecimiento que puede cambiar. Aquí hay algunos servicios con opciones gratuitas que vale la pena explorar.

1. Piktochart.com
2. Canva.com
3. Infogr.am
4. Visme
5. Easel.ly

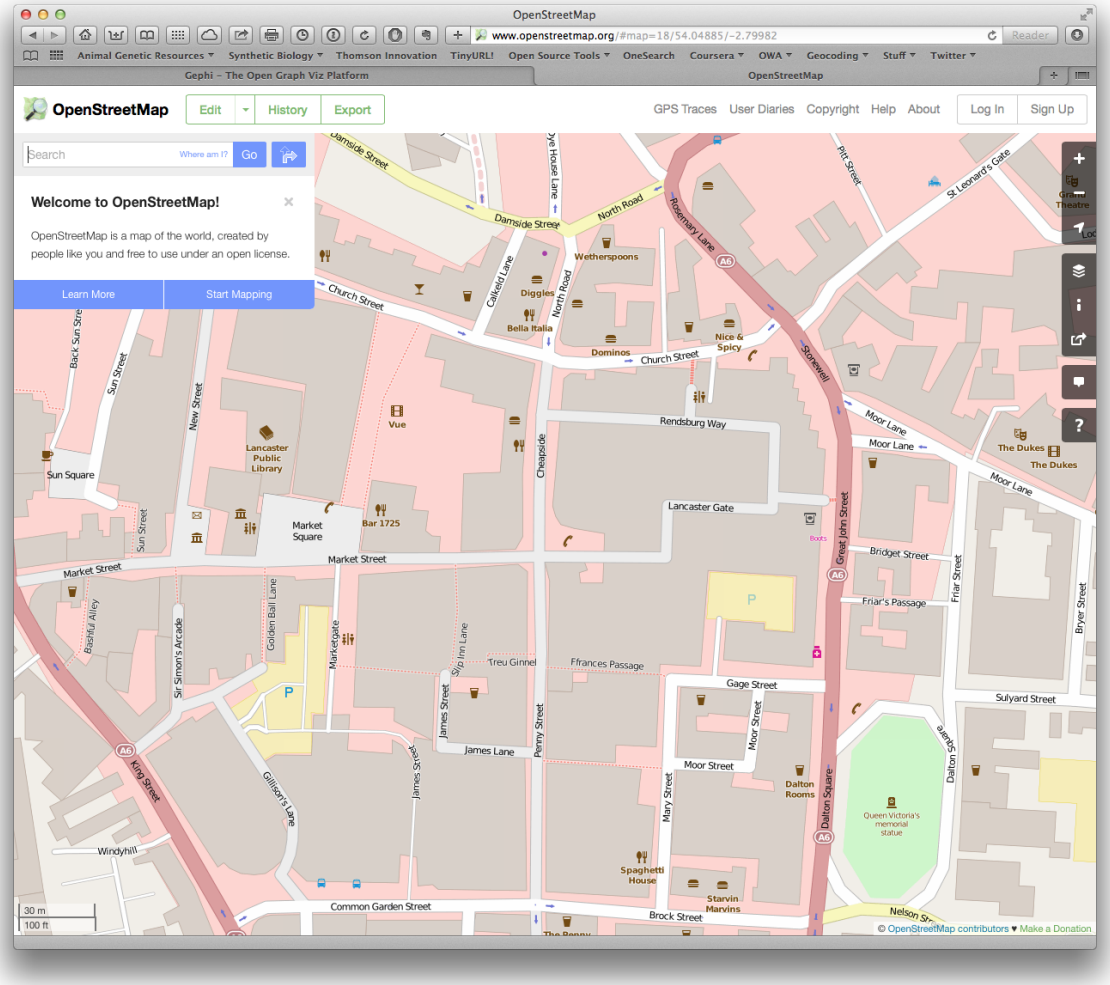
Los sitios web como Cool Infographics pueden ser útiles para encontrar fuentes adicionales, explorar lo que está de moda en el mundo de la infografía y los tutoriales. Las herramientas como Apple Keynote, Open Office Presentation o Powerpoint pueden ser muy útiles para la infografía de enmarcar (dibujar) para ver qué funciona.

2.7 Mapeo Geográfico

Además del ubicuo Google Maps o bien conocido Google Earth , creemos que vale la pena echarle un vistazo a otros servicios.

2.7.1 OpenStreetMap

Con razón popular



2.7.2 Leaflet

Una muy popular biblioteca de código abierto de JavaScript para mapas interactivos.

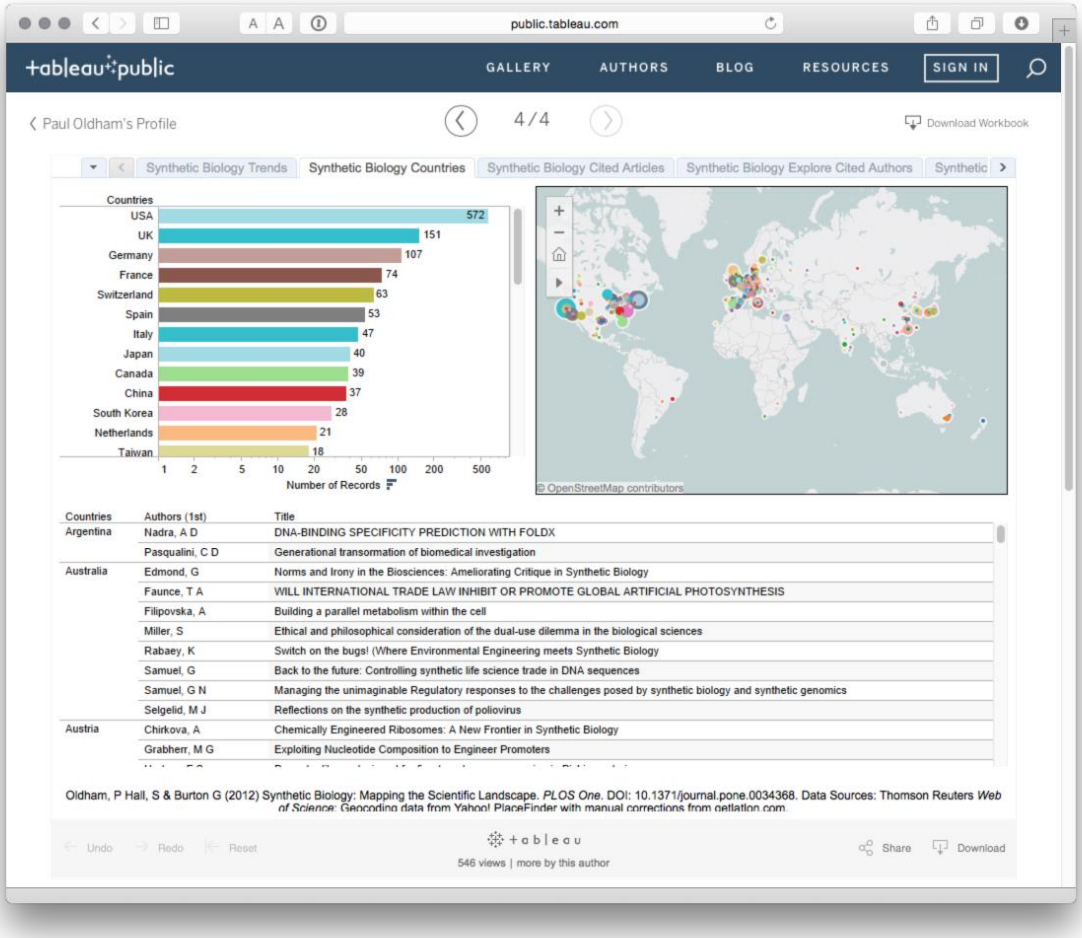
Accesible a través de una API. Los usuarios de R podrían usar el leafletr paquete con tutoriales y recorridos disponibles en R-bloggers . Para usuarios de Python intente folium aquí o aquí .

2.7.3 Tableau Public

Ya mencionado anteriormente. Tableau Public utiliza Open Street Map para crear una **poderosa combinación de gráficos interactivos** que se pueden vincular a

mapas geocodificados en varios niveles de detalle. Vea un ejemplo aquí para la literatura científica sobre biología sintética.

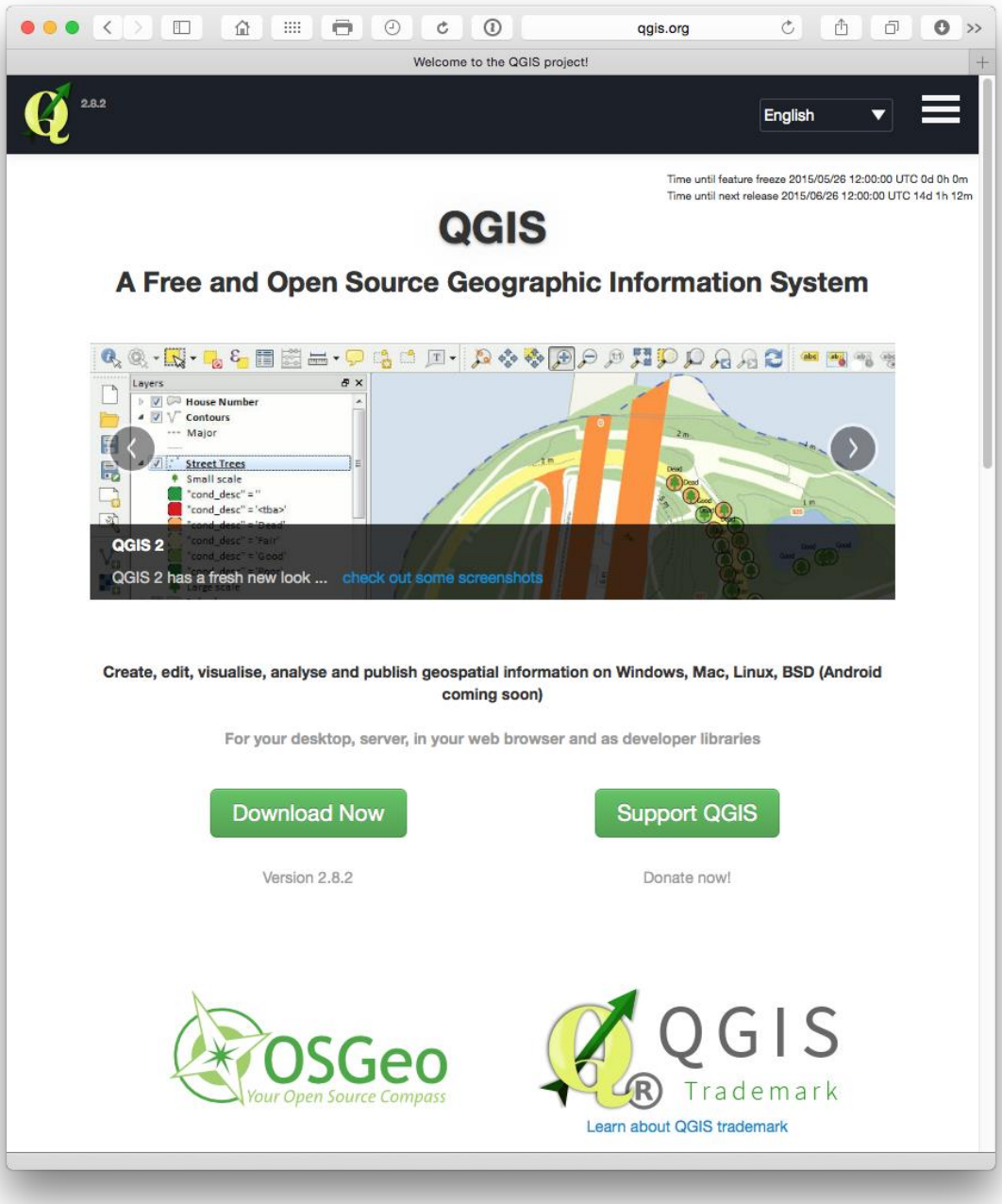
Tableau Public es probablemente la forma más fácil de comenzar a crear sus propios mapas con datos de patentes. El siguiente mapa se produjo utilizando geocodificación personalizada y conectando los datos al país de publicación y los títulos de las publicaciones científicas.



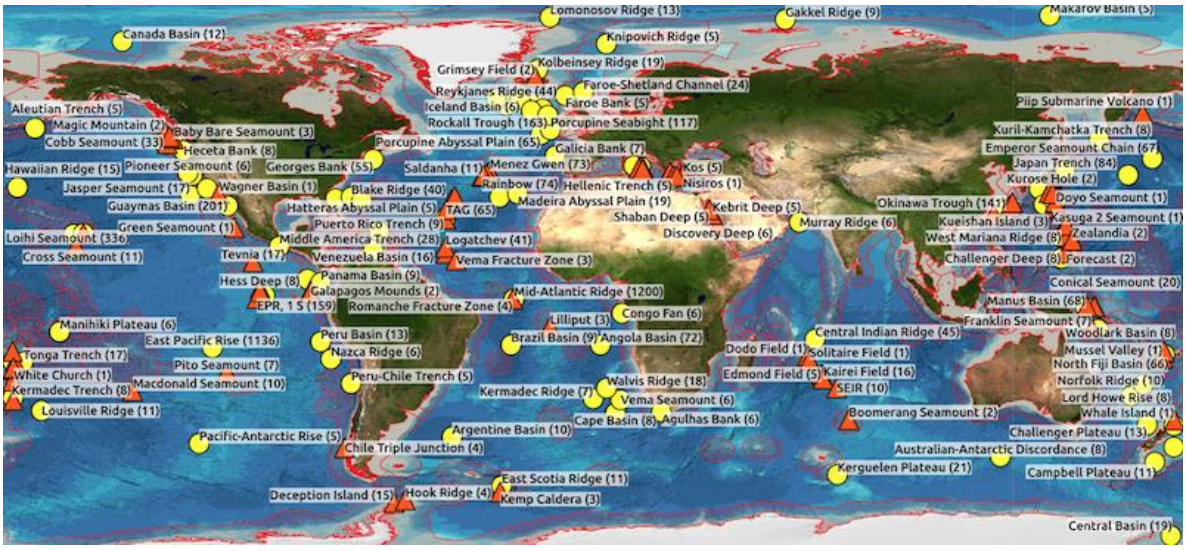
Para ver la versión interactiva prueba esta página . Es posible crear fácilmente mapas simples pero efectivos en Tableau Public.

2.7.4 QGIS

Un paquete de software muy popular y sofisticado que se ejecuta en todas las plataformas principales.

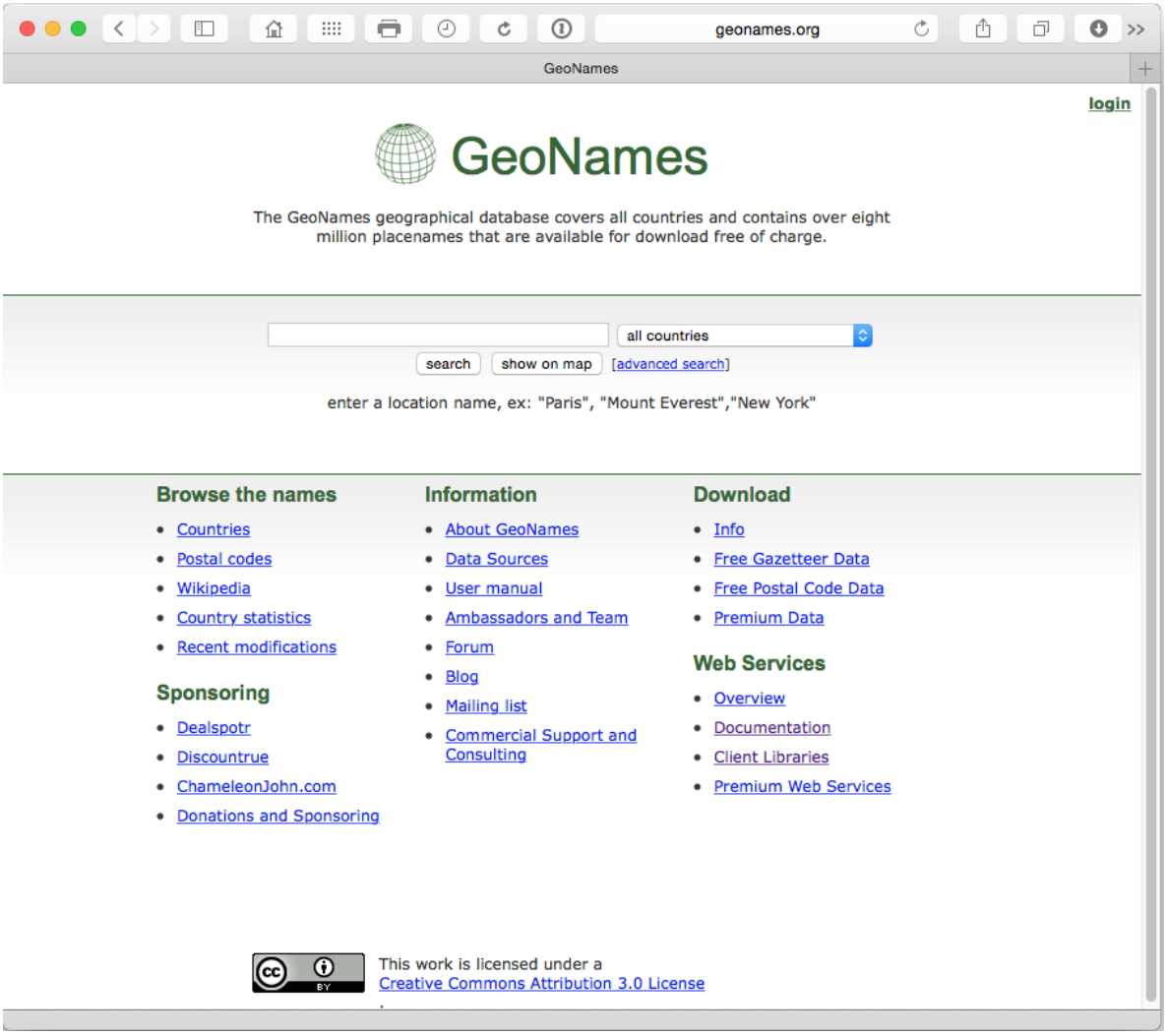


El uso de QGIS Oldham y Hall et al. Cartografiaron las ubicaciones geográficas mundiales de la investigación científica marina y los documentos de patentes que hacen referencia a las ubicaciones de aguas profundas, como los respiraderos hidrotermales (consulte la sección Valoración de las profundidades). Este es un mapa de baja resolución QGIS de lugares de investigación científica en los océanos basado en la minería de textos de la literatura científica.



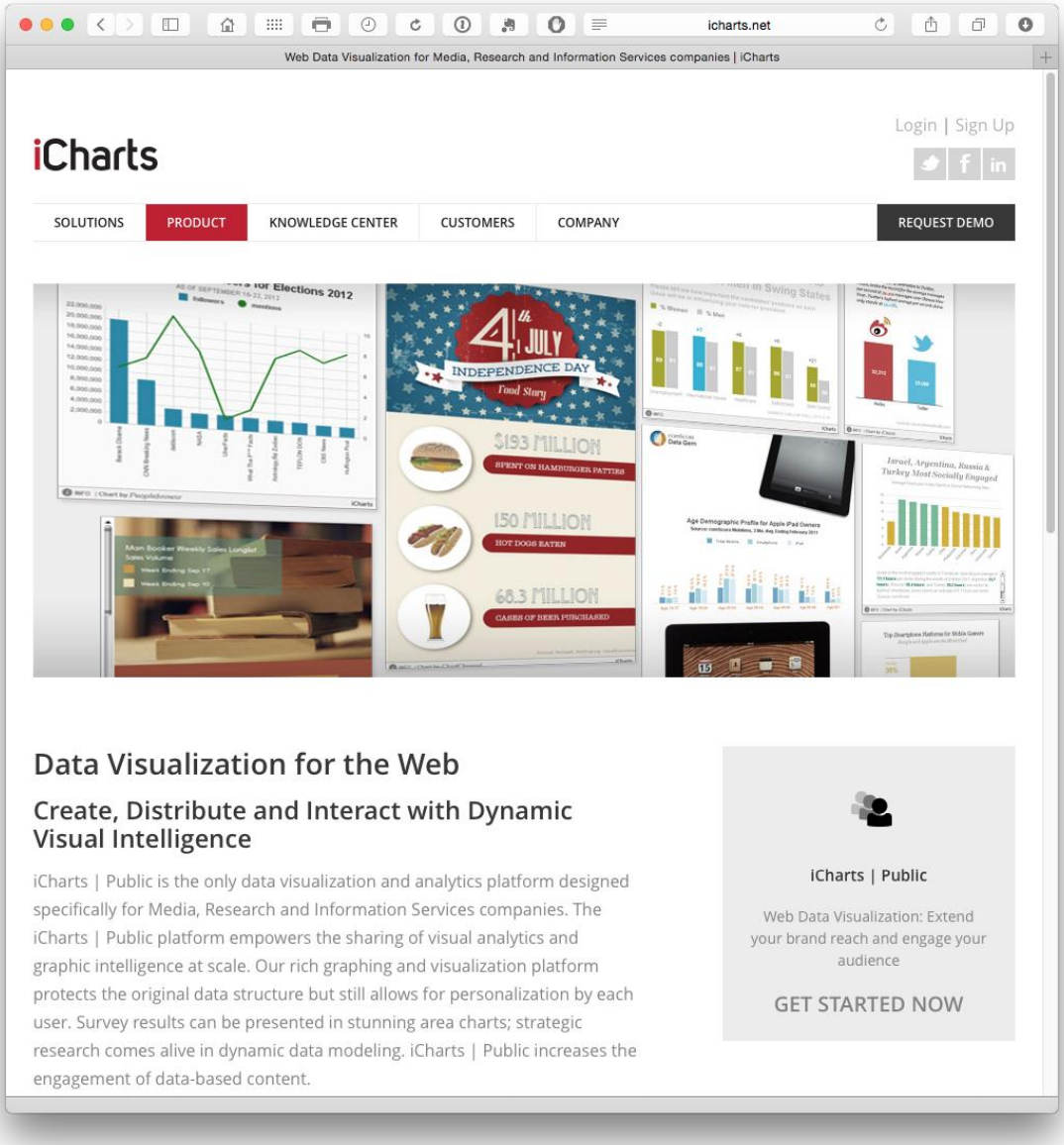
2.7.5 Geonames.org .

No es un programa de mapeo, en cambio, geonames es una base de datos increíblemente útil de nombres de lugares georreferenciados de todo el mundo junto con un servicio web RESTful . Si necesita obtener los datos georreferenciados para un gran número de lugares, esta debe ser su primera parada. Se puede acceder a los geonames en R usando las geonames bibliotecas de cliente junto con Python, Ruby, PHP y otros.



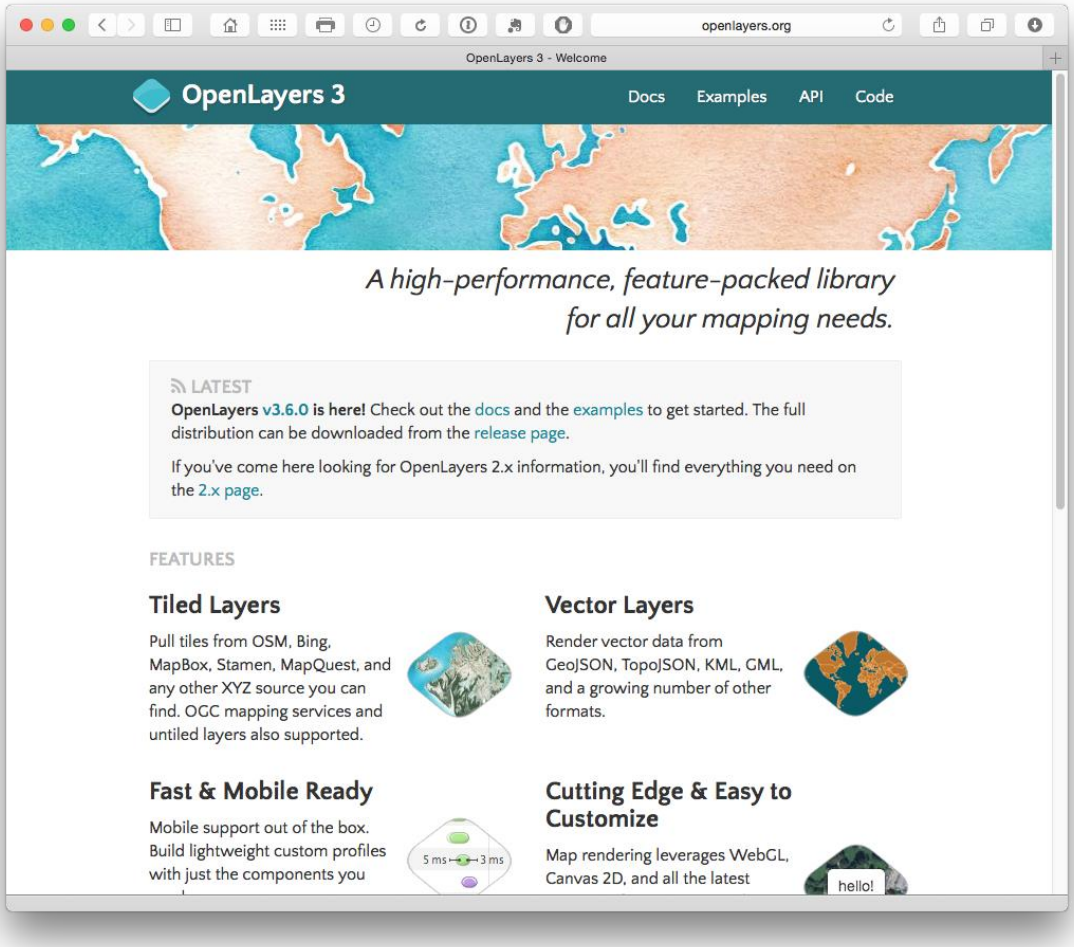
2.7.6 iCharts

Un servicio de visualización de datos gratuito y premium:



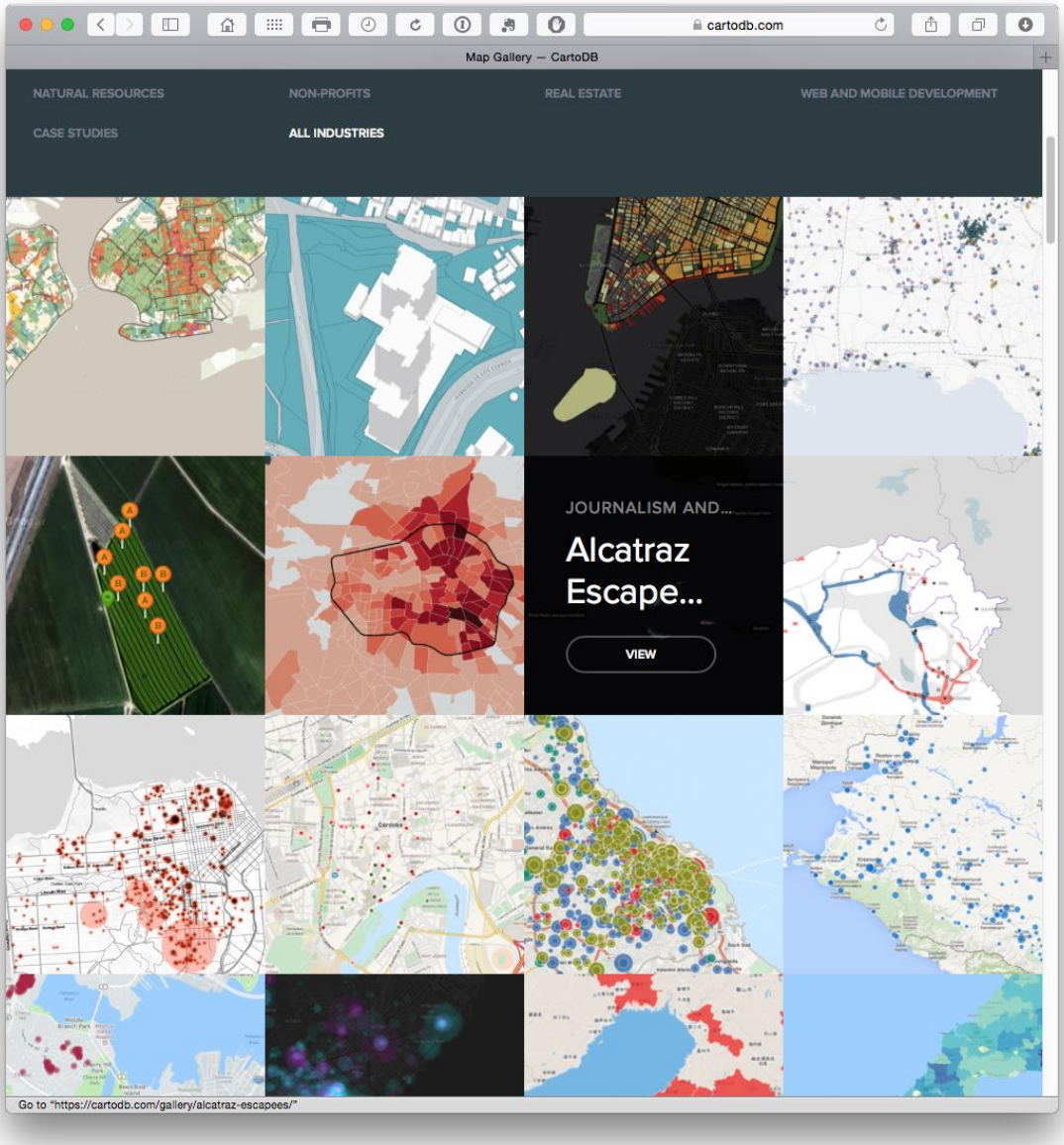
2.7.7 OpenLayers3

OpenLayers3 le permite agregar sus propias capas a OpenStreetMap y otras fuentes de datos y puede resultar muy útil si está buscando crear sus propias capas. También tiene una API y tutoriales.



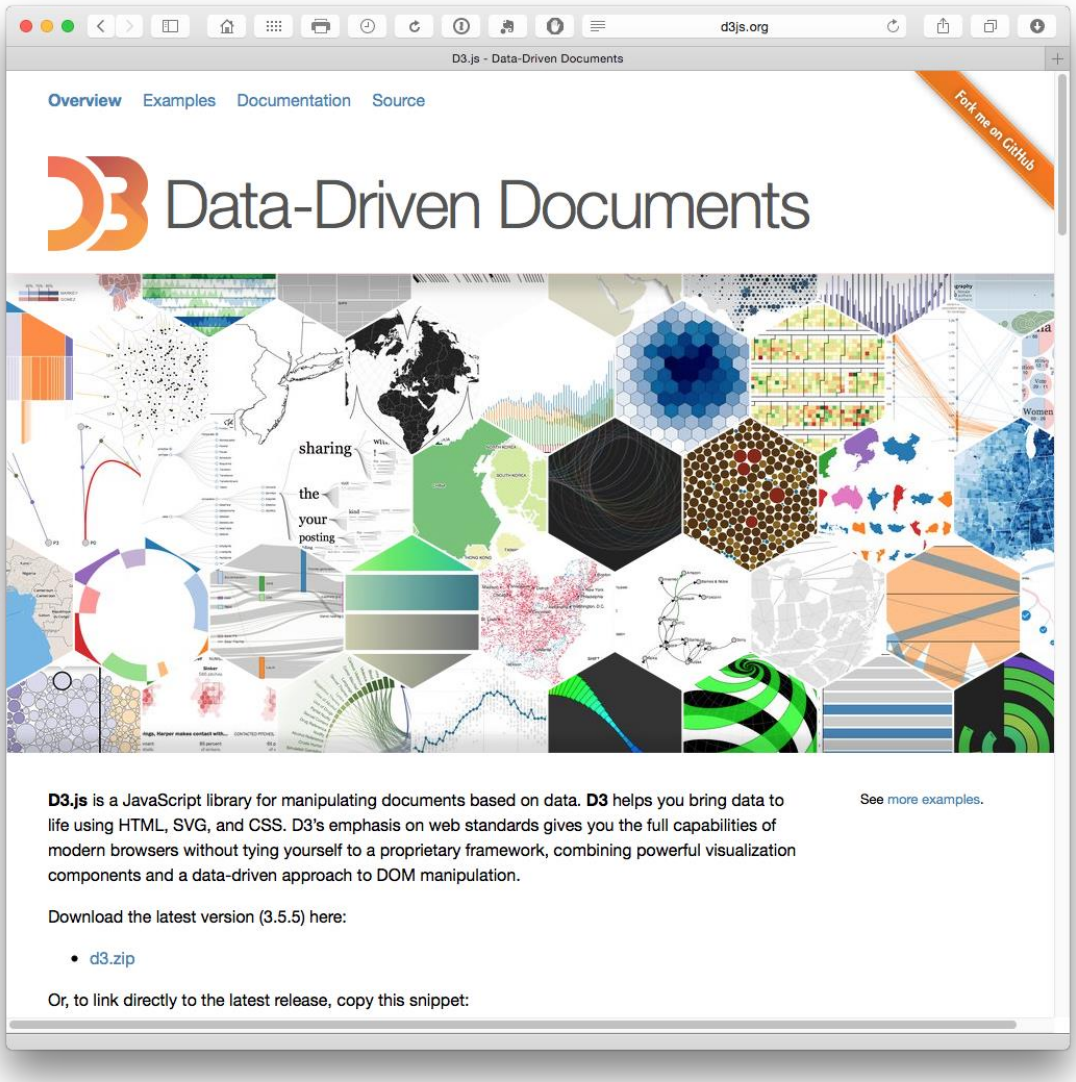
2.7.8 CartoDB

Cuentas gratuitas y de pago con una bonita galería de ejemplos.



2.7.9 d3.js

Una biblioteca de javascript para manipular datos y documentos. Esta es la biblioteca detrás de algunas de las otras herramientas de visualización mencionadas con frecuencia en la web.

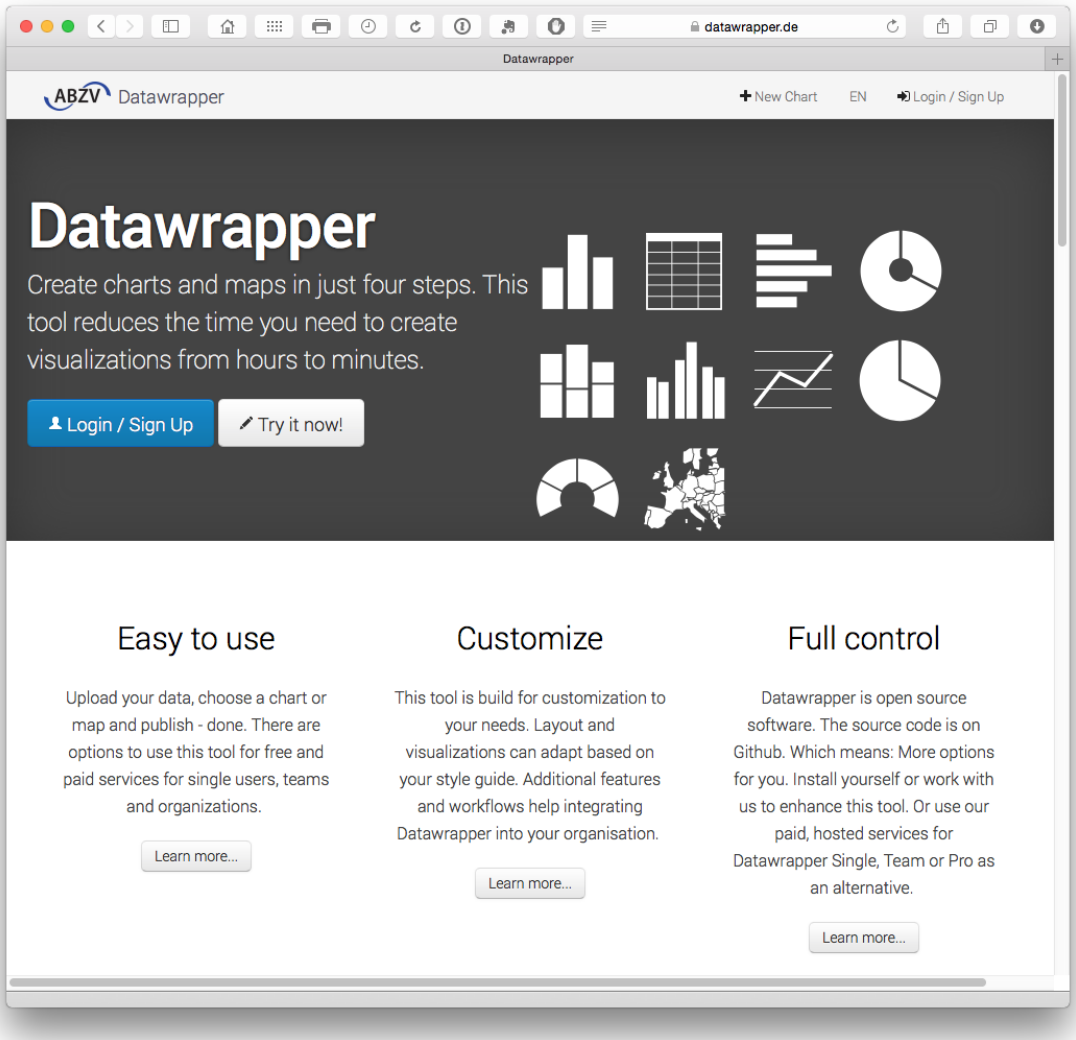


2.7.10 Highcharts

Gratis para uso no comercial con una variedad de planes de precios.

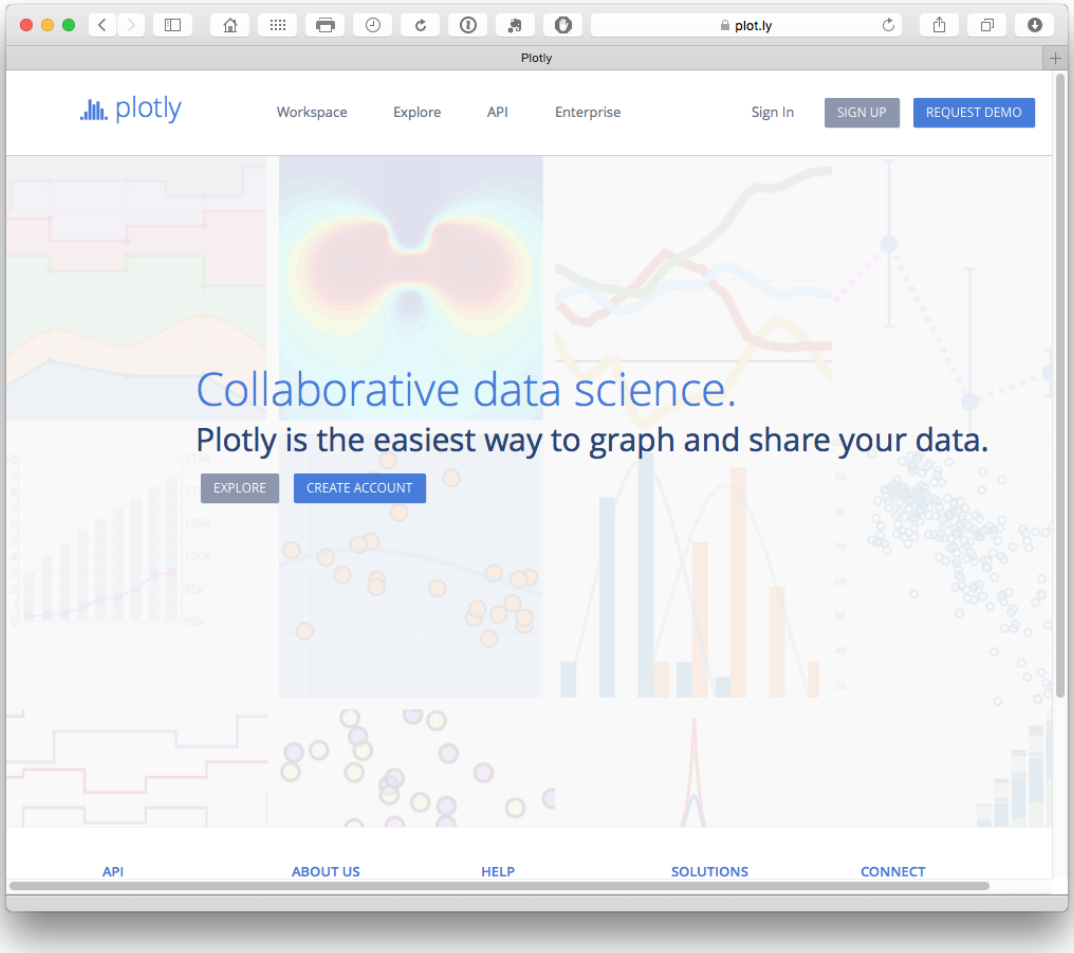
2.7.11 [Datawrapper]

Un servicio totalmente de código abierto para crear gráficos y mapas con sus datos. Ampliamente utilizado por los grandes periódicos, por lo que los gráficos serán familiares. Cree una cuenta o bifurque la fuente de Github aquí . Hay una opción gratuita y un conjunto de planes de precios.



2.7.12 Plotly

Gratis y con una API con clientes para R, Python y Matlab, Plotly es un servicio gratuito cada vez más popular que utiliza la biblioteca D3.js mencionada anteriormente con la versión empresarial utilizada por compañías como Google. Plotly es cada vez más popular y tiene una gama de clientes API para Python, Matlab, R, Node.js y Excel. La facilidad de uso y acceso de Plotly desde una variedad de entornos son grandes razones para su éxito creciente.

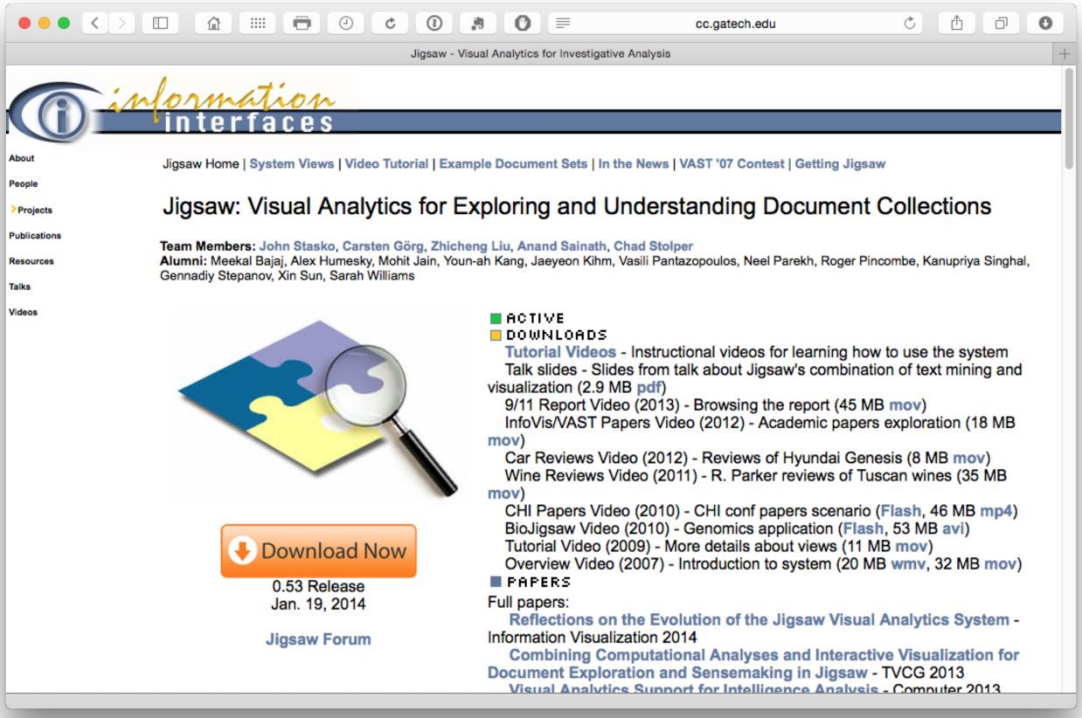


2.8 Minería de textos

Hay muchas herramientas de minería de texto y muchas de ellas son gratuitas o de código abierto. Éstos son algunos de los que hemos encontrado.

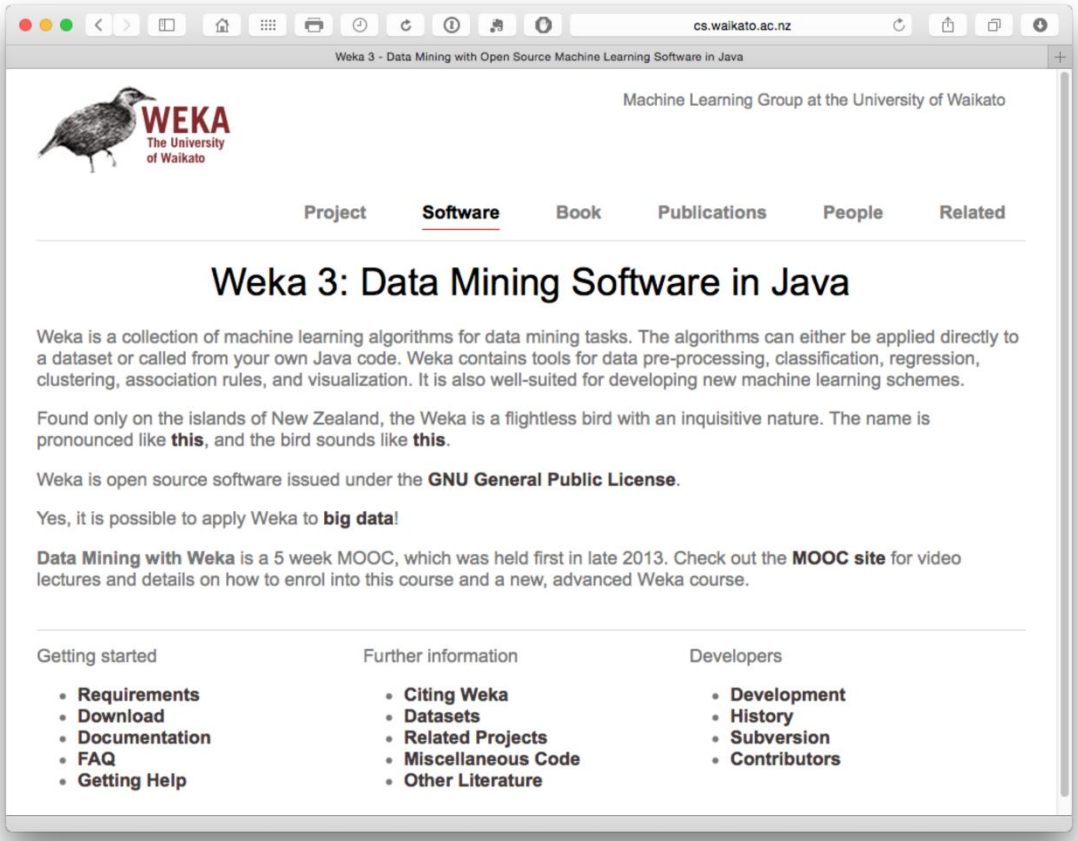
2.8.1 Jigsaw Visual Analytics

Para explorar y comprender colecciones de documentos.



2.8.2 Weka

Software de minería de texto basado en Java.



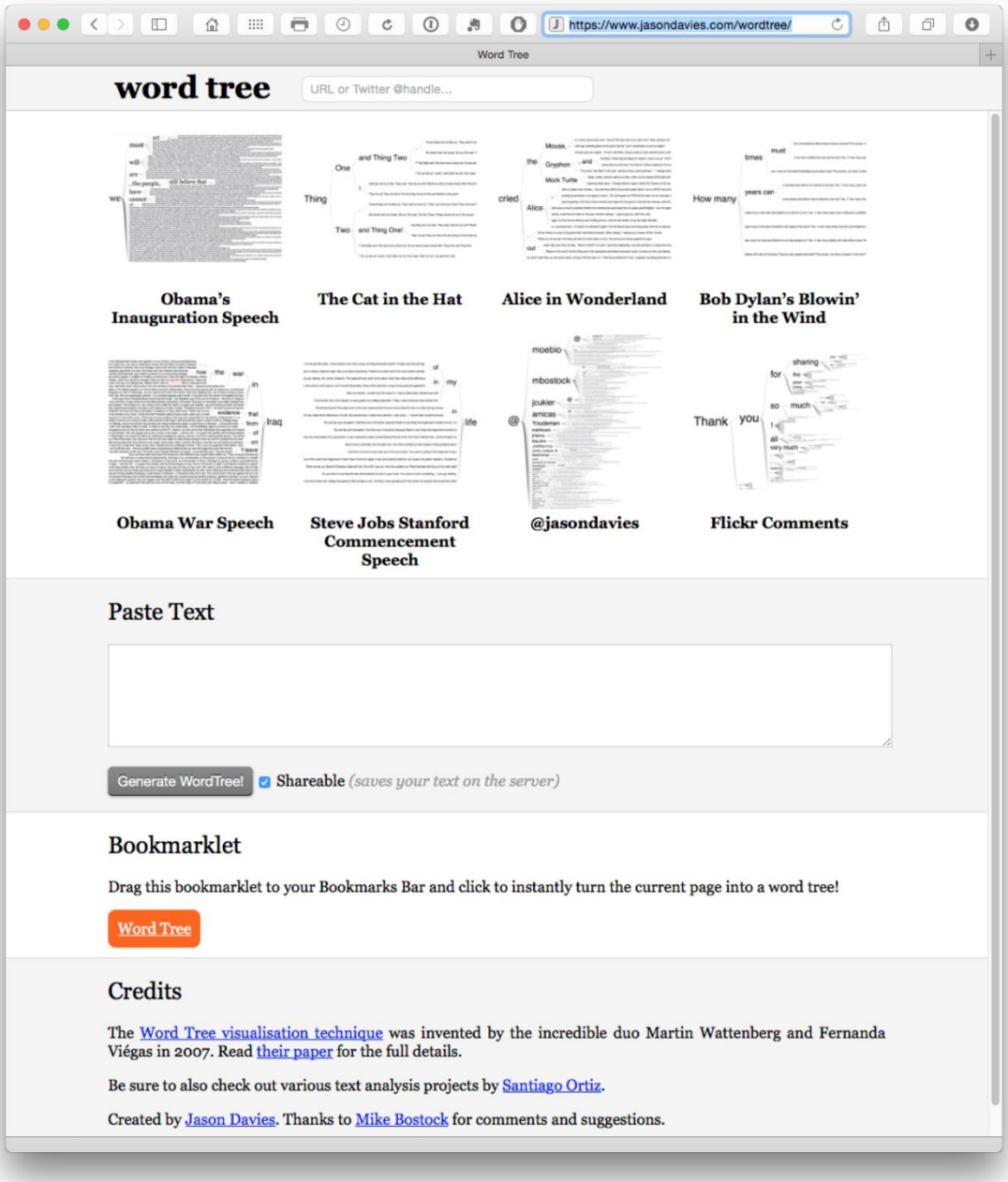
2.8.3 Árboles de palabras

Los árboles de palabras se pueden utilizar para la investigación detallada de textos como los árboles de reclamaciones. Los dos primeros ejemplos están tomados de las Directrices de la OMPI para la preparación de informes de patentes .

2.8.4 Los árboles de Google Word

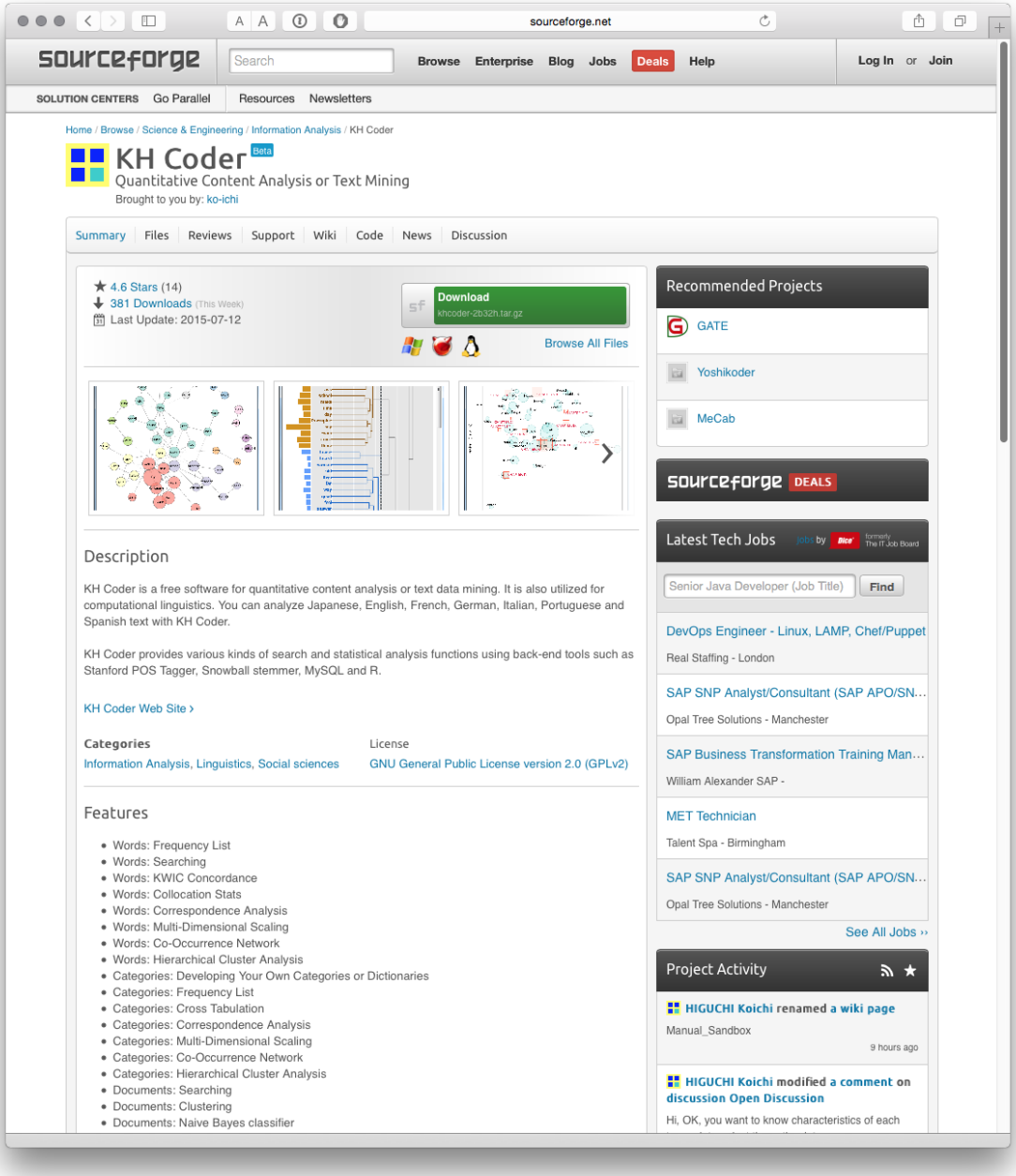
En el sitio de desarrolladores de Google proporcionan instrucciones para generar árboles de palabras con Javascript.

y [Jason Davies](#) , creador de árboles.



2.8.5 KH Coder

Software libre que permite el análisis de contenido cuantitativo / minería de texto.



2.8.6 R y el tmpaquete

El tmpaquete en R (por ejemplo, utilizando RStudio) proporciona acceso a una gama de herramientas de minería de texto. Para una introducción de los desarrolladores de paquetes vea aquí . Una serie de tutoriales muy útiles también están disponibles para la extracción de texto en R-bloggers . Para un enfoque paso a paso, vea Graham Williams (2014) Hands-On Data Science con R Text Mining .

Para obtener una descripción general reciente de las herramientas de minería de texto en R, vea la Vista de Tarea CRAN de Fridolin Wild (2014): Procesamiento en lenguaje natural que enumera los distintos paquetes y sus usos.

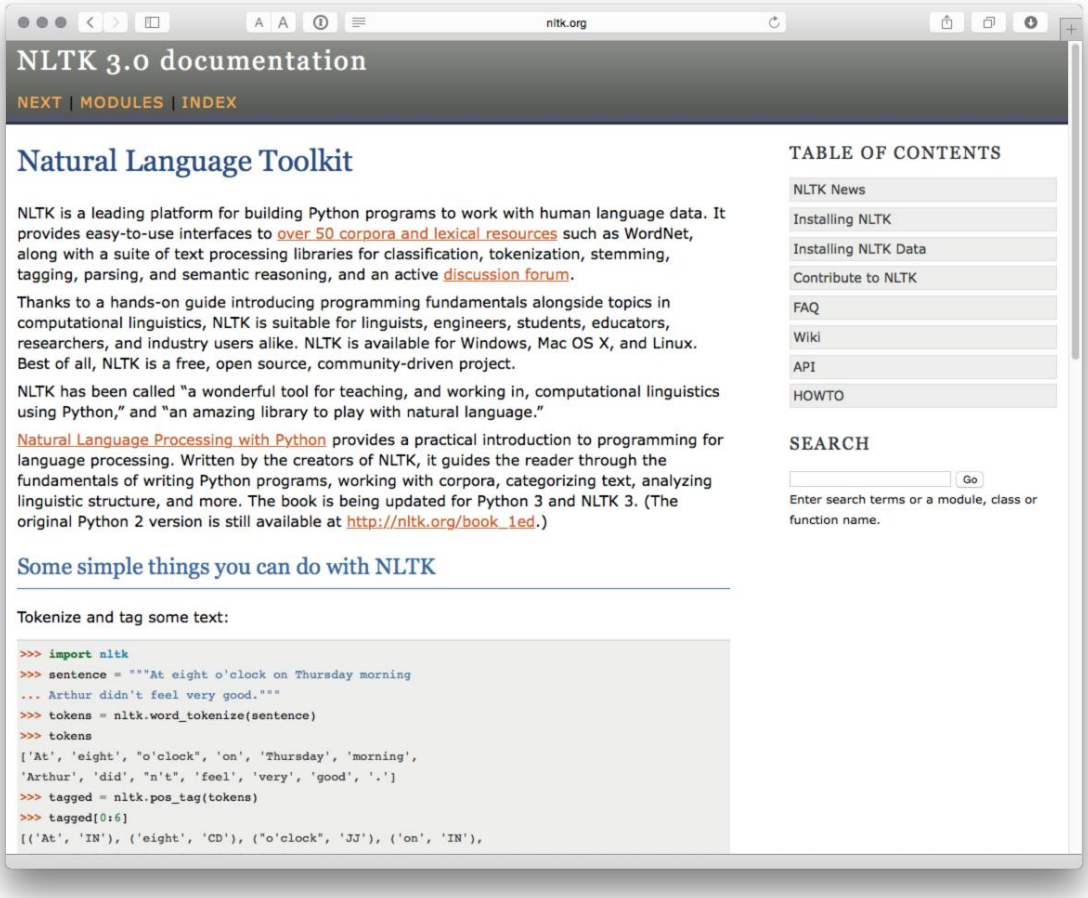
Tenga en cuenta que muchos paquetes de minería de texto en general se enfocan en generar palabras. Para fines no académicos esto no es muy útil. El análisis de patentes típicamente se enfocará en extraer y analizar fraces (ngrams). Por lo tanto, busque herramientas que extraigan frases y permitan que sean interrogadas en profundidad.

2.8.7 Python y minería de texto

Hay bastantes recursos disponibles en minería de texto usando Python. Tenga en cuenta que Python puede estar por delante de R en términos de recursos de minería de texto (hasta que se compruebe que estamos equivocados). Sin embargo, tenga en cuenta que Python y R se utilizan cada vez más juntos para explotar sus diferentes fortalezas. Aquí hay algunos recursos para ayudarlo a comenzar.

2.8.8 El kit de herramientas de lenguaje natural (NTLK)

NTLK parece ser el paquete líder y cubre casi todas las necesidades principales. El libro que acompaña el procesamiento del lenguaje natural con Python también puede valer la pena considerar. El paquete de minería de texto de Python es más simple que el paquete NTLK gigante pero puede satisfacer sus necesidades.



Este tutorial detallado puede ser útil para aquellos que desean comenzar con el paquete NTLK en Python.

2.8.9 Otros recursos de minería de texto

Para una gama más amplia de opciones de minería de texto, consulte este artículo de análisis predictivo sobre las 20 herramientas de software de minería de texto gratuitas más importantes .

Para otras herramientas de minería de texto gratuitas, pruebe algunos de los sitios web de lingüística de corpus como The Linguist List , esta lista o esta lista . Tenga en cuenta que la mayoría de estas herramientas están diseñadas pensando en los lingüistas y que un número considerable puede estar mostrando su edad. Sin embargo, incluso las herramientas de concordancia simples, como AntConc, pueden desempeñar un papel importante en el filtrado de grandes cantidades de documentos para extraer información útil.

Algunas herramientas de análisis como VantagePoint de Search Technology Inc. se han desarrollado y adaptado especialmente para procesar datos de patentes y están disponibles en una versión subsidiada para estudiantes de vpinstitute . También hay una serie de herramientas de software de análisis de datos cualitativos que se pueden aplicar al análisis de patentes como MAXQDA , NVivo , Atlas TI y QDA Miner . Sin embargo, con la excepción de QDA Miner Lite(solo para Windows), si bien ofrecen pruebas gratuitas, no se incluyen en la categoría de software de código abierto o gratuito que es nuestro enfoque.

2.9 Redondear

En este capítulo hemos cubierto algunas de las principales herramientas gratuitas y de código abierto que están disponibles para el análisis de patentes. Estas no son herramientas específicas de patentes, pero pueden adaptarse fácilmente al análisis de patentes. Con la excepción de la limpieza de los nombres de solicitantes e inventores de patentes y campos concatenados, los datos de patentes son muy adecuados para la visualización y el mapeo de redes. La disponibilidad de datos a nivel de país, campos de dirección y nombres de lugares en los textos también significa que los datos de patentes se pueden utilizar fácilmente para el mapeo geográfico.

En la práctica, es importante identificar un conjunto de herramientas que funcionen mejor para usted y el tipo de tareas de análisis de patentes que funcionen mejor para usted.

También es importante enfatizar que en la práctica puede usar una combinación de herramientas gratuitas y no gratuitas. Por ejemplo, el reciente panorama de patentes de la OMPI para recursos genéticos animales involucró el uso de GNU Parallel y Map Reduce para la minería de textos a gran escala de 11 millones de patentes de texto completo que utilizan patrones de Ruby, combinado con el uso de PATSTAT para estadísticas, Thomson Innovation y VantagePoint para validación, y Tableau y Gephi para visualización. En resumen, es posible realizar casi todas las tareas de análisis de patentes, utilizando herramientas gratuitas, pero en la práctica, un ecosistema mixto de código abierto y herramientas comerciales puede producir el mejor flujo de trabajo para las tareas que realiza. Como tal, es importante pensar en las herramientas que se necesitan y en dónde apoyan y fortalecen los flujos de trabajo de análisis existentes.

2.10 La lista de verificación

Si se pasa al software de código abierto por primera vez, puede ser útil desarrollar una lista de preguntas básicas para evaluar si una herramienta o un conjunto de herramientas satisfarán sus necesidades particulares. La siguiente lista no pretende ser definitiva ni exhaustiva, sino que tiene como objetivo fomentar el pensamiento sobre sus requisitos particulares.

1. ¿Tiene sentido esta herramienta? Es decir, ¿está claro de inmediato cuál es el propósito de una herramienta? Si la respuesta es sí, esta es una buena señal. Si la respuesta es no, la herramienta puede ser demasiado especializada para sus necesidades particulares o los creadores pueden tener dificultades para expresar claramente lo que la herramienta está tratando de hacer (una mala señal).
2. ¿Entiendes el lenguaje en el que está escrita la herramienta? ¿Es un problema si no lo hace (ver más abajo)? ¿Vale la pena entrenar a alguien en este idioma? ¿Hay cursos gratuitos o asequibles disponibles?
3. ¿El código fuente está abierto o es propietario y cuáles son los términos y condiciones de la licencia de código abierto? Al utilizar software de código abierto o gratuito, es importante tener claro qué significan las disposiciones precisas de la licencia. Por ejemplo, ¿está obligado a realizar modificaciones en el código fuente a disposición de otros en los mismos términos que la licencia original? Si está trabajando con un código fuente, esta es una pregunta importante de IP. Si no está trabajando a nivel de código fuente, esto puede no ser un problema, pero siempre tiene sentido entender la licencia de código abierto.
4. ¿Quién posee los datos? Si carga datos en un servicio basado en web, ¿quién es el propietario de los datos una vez que se cargan y quién más puede tener acceso a ellos y bajo qué condiciones? Estas preguntas son particularmente pertinentes cuando los datos son comercialmente relevantes.
5. ¿Qué significa realmente libre? Las versiones gratuitas son a menudo una ventaja en servicios premium (de ahí el término freemium). Esta transición es una característica clave de los modelos de negocio de código abierto. En algunos casos, la libertad puede ser altamente restringida en términos de la cantidad de datos que se pueden procesar, guardar o exportar. En otros casos, no se imponen restricciones de uso de la herramienta. Sin embargo, el conocimiento sobre el uso de la herramienta puede ser la prima real o el factor de costo, especialmente si depende de esa herramienta. Esté preparado para esto.
6. ¿Qué otras compañías (u oficinas de patentes) están usando esta herramienta? Esto puede ser un indicador de confianza y también proporciona ejemplos de casos de uso concretos.

7. ¿La herramienta está bien soportada con documentación y tutoriales? Este es un indicador de madurez y "compra" por parte de una comunidad de desarrolladores y usuarios.
8. ¿Qué tan grande es la comunidad de usuarios? ¿Están activos en la creación de foros y blogs, etc. para apoyar a la comunidad más amplia de usuarios?
9. ¿Es esta una herramienta de una función o una herramienta multiuso? Es decir, ¿esta herramienta cubrirá casi todas las necesidades o es un good to have componente específico en un kit de herramientas? En algunos casos, una herramienta que hace una cosa muy bien es un activo real donde otras herramientas se caen porque intentan hacer demasiadas cosas y hacerlas mal. De las herramientas enumeradas anteriormente, R y Python (posiblemente en combinación) se acercan más a las herramientas que podrían usarse para un flujo de trabajo completo de análisis de patentes desde la adquisición de datos hasta la visualización. En la práctica, la mayoría de los kits de herramientas de análisis de patentes constarán de herramientas generales y específicas.
10. ¿Puedo romper esta herramienta? Siempre es una muy buena idea averiguar cuáles son las limitaciones del software para que no lo tomen por sorpresa cuando intente hacer algo que es de misión crítica. En particular, el software puede afirmar que realiza tareas particulares, como el manejo de miles o millones de registros, pero las hace muy mal, en todo caso. Al empujar una herramienta más allá de sus límites, es posible determinar dónde están los límites y cómo obtener lo mejor de ella.
11. ¿Es la herramienta proporcional a mis necesidades? Recientemente ha habido mucha emoción Big Data y el uso de Hadoop para tratar grandes volúmenes de datos utilizando computación distribuida. Si bien Hadoop es de código abierto, para que cualquiera pueda usarlo, su adopción generalmente sería desproporcionada para las necesidades de la mayoría de los análisis de patentes, excepto cuando se trata de casi todo el conjunto de documentos de patentes globales, grandes volúmenes de publicaciones y cantidades considerables de datos científicos. A modo de ilustración, como se señaló anteriormente, el informe de la OMPI sobre recursos genéticos animales utilizó el GNU Parallel para procesar 11 millones de registros de patentes. La decisión de usar GNU Parallel se tomó en parte sobre la base de que Hadoop habría sido complicado de implementar y exagerar para el caso de uso particular. En resumen, vale la pena considerar cuidadosamente si una herramienta es apropiada y proporcional a la tarea en cuestión.
12. Finalmente, la regla de oro para adoptar cualquier herramienta para el análisis de patentes se puede expresar en términos muy simples. ¿Esto funciona para mí?

Si tiene alguna sugerencia de herramientas gratuitas o de código abierto que deberíamos incluir en el manual, no dude en agregar un comentario a la versión electrónica de este capítulo.

2.11 créditos

El desarrollo de la lista de herramientas de código abierto se benefició de los siguientes artículos.

1. Creative Bloq 11/11/2014 Las 37 mejores herramientas para la visualización de datos
2. Nismith Sharma 2015 Las 14 mejores herramientas de visualización de datos. Noticias de TNW

Capítulo 3 Campos de datos

Este capítulo proporciona un recorrido por los campos de datos de patentes para aquellos que son completamente nuevos en el análisis de patentes o que desean comprender un poco mejor el funcionamiento de los datos de patentes. Una versión en video del recorrido está disponible aquí y la plataforma de diapositivas está disponible para descargar en .pdf , powerpoint y apple keynote de GitHub . **Este capítulo profundiza cada campo de datos y su uso en el análisis de patentes.**

3.1 ¿Qué es una patente?

Una patente se puede describir de dos maneras principales:

1. Como forma de derecho de propiedad intelectual.
2. Como un tipo de documento.

Comprender la estructura de los documentos de patentes y los campos de datos es la base esencial del análisis de patentes. Sin embargo, para aquellos que son nuevos en el sistema de patentes, vale la pena destacar las características clave de las patentes como una forma de derecho de propiedad intelectual.

3.2 Como forma de derecho de propiedad intelectual.

1. Una patente es una concesión temporal de un derecho exclusivo al titular de una patente para evitar que otras personas realicen, utilicen, ofrezcan para la venta o importen una invención patentada sin su consentimiento, en un país donde la patente esté en vigor.
2. Los derechos de patente son derechos territoriales, solo son válidos en el territorio del país donde se otorgan.
3. Las patentes generalmente se otorgan por un período de 20 años a partir de los datos de presentación de una solicitud, pero pueden ser rechazadas o revocadas.
4. Para ser elegible una **invención reivindicada** debe:
    - Involucrar materia patentable.
    - Ser nuevo o novedoso
    - Involucrar un paso inventivo.
    - Ser susceptible a la aplicación industrial o útil.

3.3 Las patentes como tipo de documento.

Para el **análisis de patentes** debemos concentrarnos en las patentes como una forma de documento y **comprender**:

1. La **estructura** de los documentos de patente **y** sus **campos de datos**.
2. Las fortalezas y limitaciones de las diferentes **bases de datos de patentes** como medio para obtener datos de patentes.

En este capítulo tratamos los aspectos básicos de los documentos de patentes y sus campos de datos.

3.4 Tipos de datos básicos

Al realizar el análisis de patentes, estamos tratando con **datos de siete tipos** diferentes:

1. **Fechas** (prioridad, fechas de solicitud y publicación)
2. **Números** (número de prioridad, número de solicitud, número de publicación, miembros de la familia, citas)
3. **Nombres** (solicitantes, también conocidos como Asignatarios - e Inventores)
4. **Códigos de clasificación** (por ejemplo, Clasificación Internacional de Patentes / Clasificación Cooperativa de Patentes)
5. **Campos de texto** (Título, Resumen, Descripción, Reclamaciones, Datos de secuencia)
6. **Imágenes** (Diagramas)
7. **Información adicional** (estado legal, registro público, etc.)

Caminaremos a través de cada uno de estos campos utilizando una solicitud de patente para genomas sintéticos del Instituto J. Craig Venter como ejemplo. En la versión electrónica, cada uno de los títulos de las imágenes está hipervinculado a sus fuentes para facilitar la exploración de los datos a medida que los recorre.

3.4.1 genomas sintéticos

La biología sintética (y la genómica sintética) comenzaron a aparecer en los titulares internacionales con la noticia en 2010 de que los miembros del Instituto J. Craig Venter habían sintetizado con éxito el genoma de un microbio de Mycoides y habían trasplantado el genoma en la celda vacía de otro Mycoides booted up. Esto llevó a un considerable entusiasmo por la creación de vida artificial y es parte de la historia de la creciente prominencia de la biología sintética. Para

nuestros propósitos, es un ejemplo interesante para recorrer los campos de datos de patentes estándar.

3.4.2 Página principal original

Lo que vemos a continuación es la portada de una solicitud del Tratado de Cooperación de Patentes (PCT) internacional del Instituto J. Craig Venter sobre Genomas Sintéticos. **El PCT permite a los solicitantes presentar una solicitud única para una posible consideración en hasta otros 148 países que son Partes en el PCT** según las decisiones tomadas por los solicitantes y las decisiones de examen en países y regiones individuales. La página principal (o biblio) muestra los campos de datos que normalmente se utilizan en el análisis de patentes.

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
28 February 2008 (28.02.2008)

PCT

(10) International Publication Number
WO 2008/024129 A2

(51) International Patent Classification:
C07H 21/04 (2006.01) *C12P 1/04* (2006.01)
C12N 5/06 (2006.01)

(21) International Application Number:
PCT/US2006/046803

(22) International Filing Date:
6 December 2006 (06.12.2006)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/742,542 6 December 2005 (06.12.2005) US

(71) Applicant *(for all designated States except US)*: **J. CRAIG VENTER INSTITUTE** [US/US]; 9704 Medical Center Drive, Rockville, MD 20850 (US).

(71) Applicant *(for US only)*: **HUTCHISON, Clyde, A., III** [US/US]; c/o J. Craig Venter Institute, 9704 Medical Center Drive, Rockville, MD 20850 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **VENTER, Craig, J.** [US/US]; c/o J. Craig Venter Institute, 9704 Medical Center Drive, Rockville, MD 20850 (US). **SMITH, Hamilton, O.** [US/US]; c/o J. Craig Venter Institute, 9704 Medical Center Drive, Rockville, MD 20850 (US).

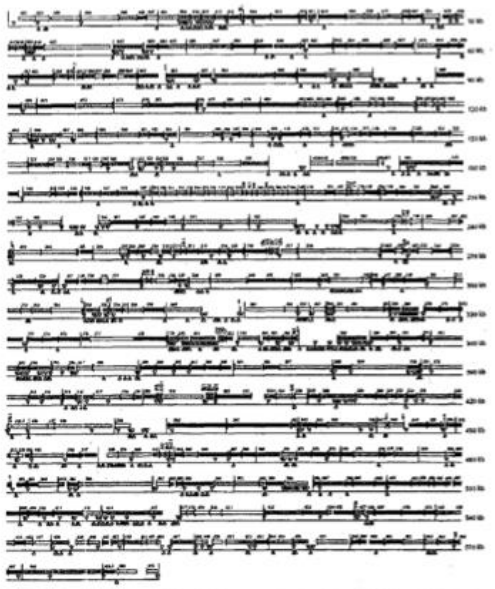
(74) Agents: **BATHURST, Brian** et al.; Carr & Ferrell LLP, 2200 Geng Road, Palo Alto, CA 94303 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LV, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, SV, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH,

[Continued on next page]

(54) Title: SYNTHETIC GENOMES



(57) Abstract: Methods are provided for constructing a synthetic genome, comprising generating and assembling nucleic acid cassettes comprising portions of the genome, wherein at least one of the nucleic acid cassettes is constructed from nucleic acid components that have been chemically synthesized, or from copies of the chemically synthesized nucleic acid components. In one embodiment, the entire synthetic genome is constructed from nucleic acid components that have been chemically synthesized, or from copies of the chemically synthesized nucleic acid components. Rational methods may be used to design the synthetic genome (e.g., to establish a minimal genome and/or to optimize the function of genes within a genome, such as by mutating or rearranging the order of the genes). Synthetic genomes of the invention may be introduced into vesicles (e.g., bacterial cells from which part or all of the resident genome has been removed, or synthetic vesicles) to generate synthetic cells. Synthetic genomes or synthetic cells may be used for a variety of purposes, including the generation of synthetic fuels, such as hydrogen or ethanol.

WO 2008/024129 A2

A partir de las fechas. **Hay tres fechas** en esta aplicación.

1. La primera fecha es el priority date(6 de diciembre de 2005) en el Priority Datacampo (30). Esto se refiere al original (primera fecha de presentación) para una solicitud en los EE. UU. Que es la prioridad (o principal) de todas las presentaciones posteriores de la misma solicitud en cualquier otro lugar del mundo (conocida como familia de patentes).
2. La segunda fecha es el International Filing Date(22) que es 12 meses después de la presentación de prioridad (US60742542).
3. La tercera fecha es el International Publication Date(campo 43) que es poco más de 24 meses después de la fecha de presentación internacional y 3 años desde la primera fecha de presentación (solicitud de prioridad).

Para el análisis de patentes, las fechas más importantes son generalmente la **fecha de prioridad** (06.12.2005) y la **fecha de publicación** (28.02.2008). La fecha de prioridad es importante por dos razones. Primero, en términos legales, establece la reivindicación de prioridad para esta invención reclamada sobre otras reclamaciones a la misma invención presentadas en el mismo período o más adelante bajo los términos del Convenio de París . Segundo, en el análisis económico, la fecha de prioridad es la fecha más cercana a la inversión en investigación y desarrollo y, por lo tanto, la más importante en el análisis económico (consulte el Manual de estadísticas de patentes de la OCDE).). Sin embargo, esta información solo está disponible cuando se publica una aplicación. Eso es típicamente 24 meses desde la fecha de presentación original. Como resultado, esta información cae desde un acantilado a medida que nos acercamos al presente.

La **fecha de publicación** es importante porque, al igual que el **número de publicación**, generalmente es la más accesible en las bases de datos de patentes. Sin embargo, en este caso hay un intervalo de 2 a 3 años entre la primera fecha de presentación y la fecha de publicación. Para el análisis de patentes, esto significa que los recuentos basados en la fecha de publicación siempre muestran tendencias que van de 2 a 3 años después de la actividad original. Sin embargo, debido a que los solicitantes deben pagar en cada etapa del proceso, los datos de publicación de patentes pueden ser útiles como un indicador de la demanda de derechos de patentes en uno o más países. En muchos casos, la fecha de publicación será la única fecha disponible para mapear tendencias.

Una lección importante de la comprensión de los campos de fecha de patente es que **los datos de patente siempre son históricos**. Es decir, siempre se refiere a la actividad en el pasado.

Abordaremos otros campos de datos a continuación. Por ahora, tenga en cuenta la información del solicitante y el inventor, incluida la dirección y otra información

útil para el análisis de patentes en la página principal (campos 71 y 72). Además, tenga en cuenta los datos de la Clasificación Internacional de Patentes como un indicador de áreas tecnológicas expresadas a través de códigos alfanuméricos (por ejemplo, C07H21 / 04 que nos dice que la invención reivindicada involucra ácidos nucleicos). También vemos campos de texto (para la minería de textos) en el título y el resumen, y finalmente tenemos una imagen con información sobre casetes de ADN que forman parte de la invención.

3.4.3 espacenet portada

Aquí podemos ver la misma información en la página principal para el registro en la base de datos espacenet. espacenet es fácilmente accesible y popular. Incluso cuando se usan herramientas comerciales, **espacenet es a menudo la forma más rápida de buscar o verificar información**. Para una breve descripción, vea estos videos y el asistente interactivo de espacenet .

Europäisches Patentamt
European Patent Office
Office européen des brevets

Espacenet
Patent search

Deutsch English Français
Contact
Change country

About Espacenet Other EPO online services

Search | Result list | My patents list (0) | Query history | Settings | Help

Refine search → Results → IL192041 (A) → Family → ... → Family → WO2008024129 (A2)

WO2008024129 (A2)
- **Bibliographic data**
- Description
- Claims
- Mosaics
- Original document
- Cited documents
- Citing documents
- INPADOC legal status
- INPADOC patent family

Quick help
- What is meant by high quality text as facsimile?
- What does A1, A2, A3 and B stand for after a European publication number?
- What happens if I click on "In my patents list"?
- What happens if I click on the "Register" button?
- Why are some sidebar options deactivated for certain documents?
- How can I bookmark this page?
- Why does a list of documents with the heading "Also published as" sometimes appear, and what are these documents?
- What is Global dossier?
- Why do I sometimes find the abstract of a corresponding document?
- What happens if I click on the red "patent translate" button?

Bibliographic data: WO2008024129 (A2) — 2008-02-28

In my patents list | Previous 9 / 9 Next | EP Register | Report data error | Print

SYNTHETIC GENOMES

Page bookmark WO2008024129 (A2) - SYNTHETIC GENOMES

Inventor(s): VENTER CRAIG J [US]; SMITH HAMILTON O [US] +

Applicant(s): CRAIG VENTER INST J [US]; HUTCHISON CLYDE A III [US]; VENTER CRAIG J [US]; SMITH HAMILTON O [US] +

Classification: - international: *C07H21/04; C12N5/06; C12P1/04*

- cooperative: C12N15/10; C12N15/1093; C12N15/66

Application number: WO2006US46803 20061206

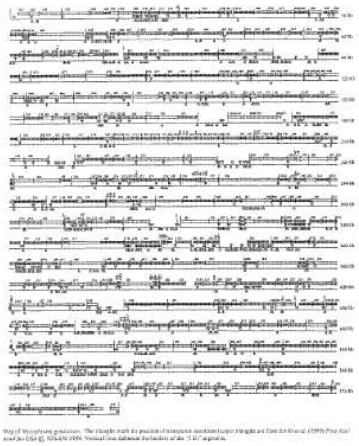
Priority number(s): US20050742542P 20051206

Also published as: WO2008024129 (A3) US2007264688 (A1) JP2009518038 (A) JP5106412 (B2) IL192041 (A) → more

Abstract of WO2008024129 (A2)

Translate this text into Albanian — patenttranslate powered by EPO and Google

Methods are provided for constructing a synthetic genome, comprising generating and assembling nucleic acid cassettes comprising portions of the genome, wherein at least one of the nucleic acid cassettes is constructed from nucleic acid components that have been chemically synthesized, or from copies of the chemically synthesized nucleic acid components. In one embodiment, the entire synthetic genome is constructed from nucleic acid components that have been chemically synthesized, or from copies of the chemically synthesized nucleic acid components. Rational methods may be used to design the synthetic genome (e.g., to establish a minimal genome and/or to optimize the function of genes within a genome, such as by mutating or rearranging the order of the genes).; Synthetic genomes of the invention may be introduced into vesicles (e.g., bacterial cells from which part or all of the resident genome has been removed, or synthetic vesicles) to generate synthetic cells. Synthetic genomes or synthetic cells may be used for a variety of purposes, including the generation of synthetic fuels, such as hydrogen or ethanol.



Sitemap Accessibility Legal notice Terms of use Last updated: 19.01.2015 Worldwide Database 5.8.22.2; 92p

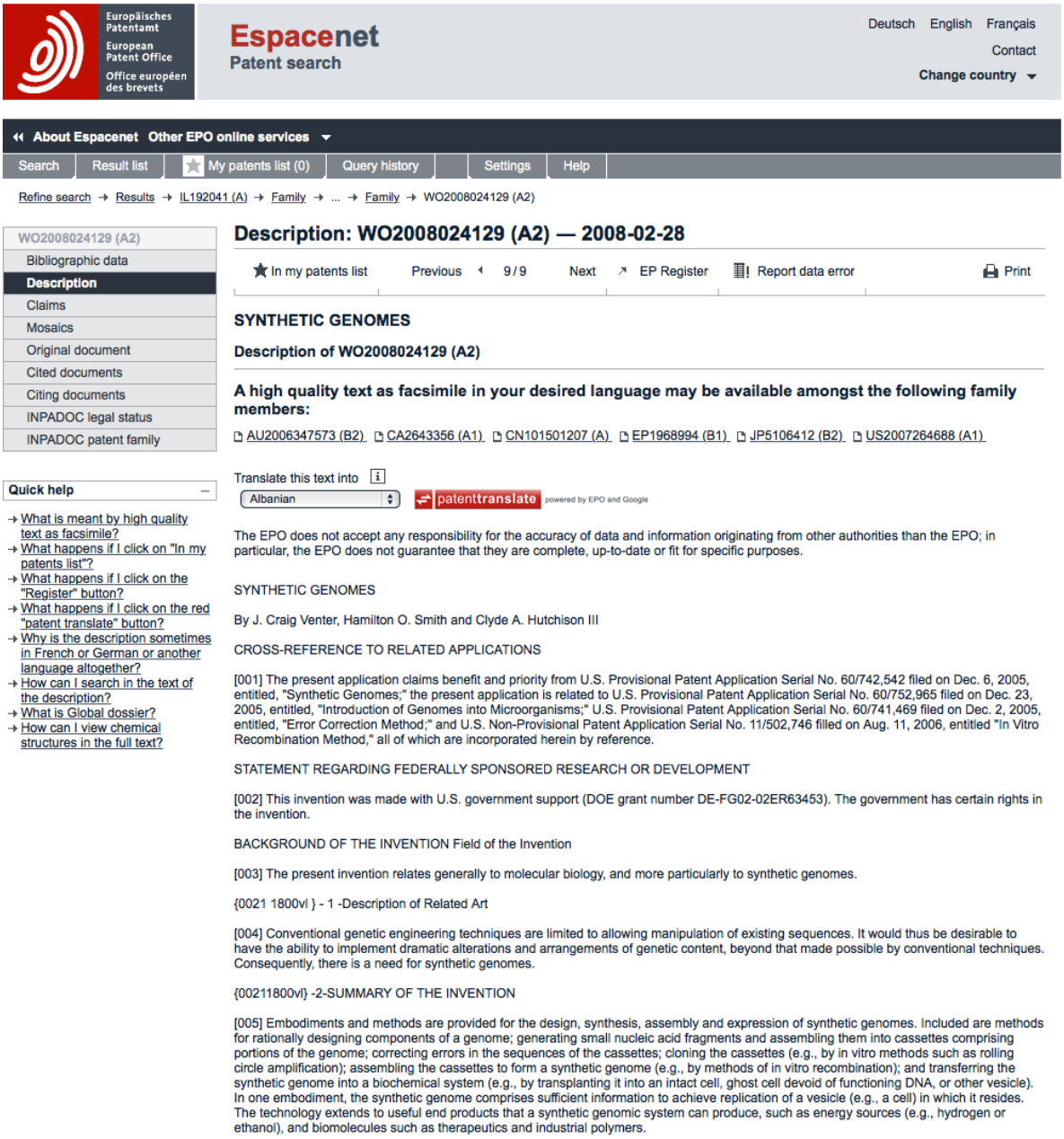
En este caso, vemos datos en el formulario que normalmente se obtiene de una base de datos de patentes como espacenet. Comenzando con las **fechas** , podemos ver que el campo del número de prioridad contiene la fecha de prioridad (primera presentación) 20051206 (como YYYYMMDD), vinculada al documento de prioridad US20050742542P, una aplicación provisional de los EE. UU. A esto le sigue el **número de solicitud** WO2006US46803 y **la fecha y el número de publicación y la fecha**. **El número de publicación**, WO2008024129A2, **es normalmente el más fácil de usar cuando se busca una base de datos de patentes**.

Otros aspectos para tener en cuenta son que **los campos de Solicitante e Inventor** incluyen información de **código de país** (por ejemplo, EE. UU.) Que utiliza códigos de país de dos letras estándar. Si bien esta información no siempre está disponible (especialmente para las solicitudes únicamente a nivel nacional), estos datos son muy útiles en el análisis de patentes para identificar colaboraciones entre países entre los inventores y los solicitantes. Sin embargo, tenga en cuenta que para uso estadístico es importante calcular el número de registros que poseen la información de este país o usar solo aquellas jurisdicciones donde se registra esta información.

3.4.4 Descripción

La sección de **descripción** (también llamada **especificación**) contiene detalles sobre:

1. Las solicitudes de **patentes anteriores y el estado de la técnica**, como la **literatura científica**.
2. en el caso de los Estados Unidos, los solicitantes incluyen información sobre si la investigación que condujo a la invención fue financiada por el gobierno, incluida la agencia de financiamiento y el número de contrato correspondiente.
3. un **resumen** seguido de **antecedentes** detallados de la invención reivindicada. Por lo general, esto incluirá ejemplos que pueden ser **ejemplos** reales o ejemplos en papel (proféticos).



Los datos proporcionados en la descripción pueden ser muy útiles cuando se aplican enfoques de minería de textos. Por ejemplo, los autores han extraído previamente en el texto millones de documentos para nombres de especies biológicas como en este artículo . En otros casos, puede ser conveniente investigar la descripción para obtener información sobre el país de origen de los materiales y el conocimiento tradicional , o explorar los usos de extractos particulares o compuestos químicos.

Sin embargo, cuando se trabaja con datos en la descripción, tenga en cuenta que a menudo es ruidoso y se requiere cuidado al construir una consulta. Por ejemplo,

una búsqueda de cerdos capturará una gran cantidad de datos sobre cerdos como animales, pero también como cerdos de juguete y dispositivos para limpiar tuberías (cerdos de tuberías). Por el contrario, las búsquedas de un nombre de país (como Senegal o Níger) pueden producir miles de resultados que no tienen nada que ver con ese país porque son parte de nombres de especies (por ejemplo, Acacia senegal o Aspergillus niger).

Por lo tanto, **es importante considerar y probar las consultas de búsqueda para campos de texto como el Título, Resumen, Descripción y Reclamaciones para evitar ser abrumado por resultados irrelevantes**.

3.4.5 Reclamaciones

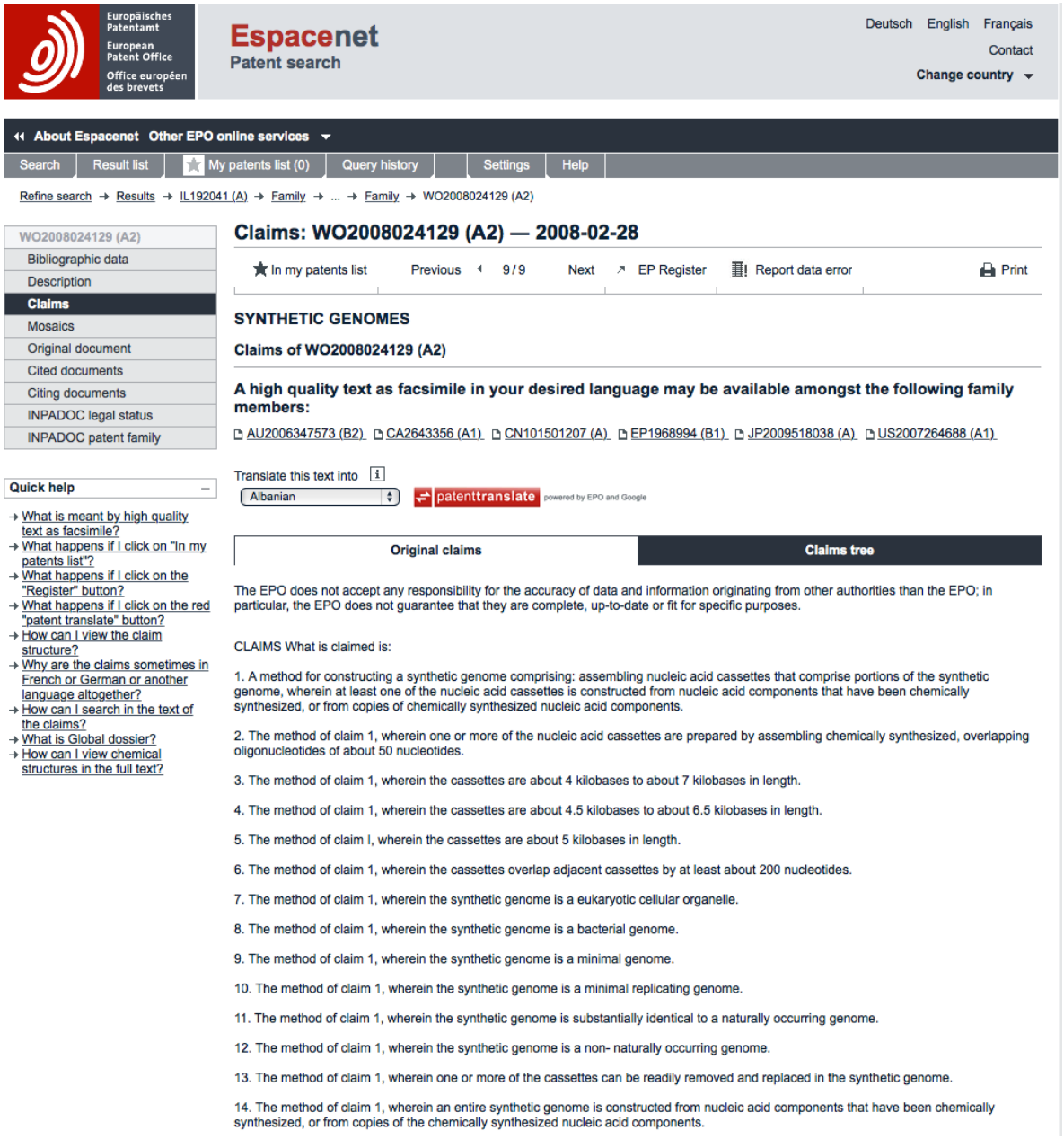
La sección de reclamaciones de un documento de patente se considera comúnmente como **la parte más importante del documento porque nos dice lo que el solicitante reclama realmente como una invención**. Lo que se reivindica en una solicitud de patente debe estar respaldado por la descripción. Por ejemplo, uno no podría insertar una sección de "Orgullo y prejuicio" de Jane Austen en la descripción de las aplicaciones de genomas sintéticos y esperar que siga adelante. Además, en países como los Estados Unidos, las solicitudes de patente se interpretan (construyen) a la luz del contenido de la descripción (consulte este artículo informativo de 2009 de Dan Burk y Mark Lemley sobre los debates en los Estados Unidos).

Las reclamaciones de patentes toman una variedad de formas y lo que puede estar permitido puede variar según el país o la jurisdicción o tomar formas especializadas (por ejemplo, patentes de diseño o patentes de plantas de EE. UU.). Eso puede dificultar la descripción e interpretación de las reivindicaciones de patentes. Para una discusión más detallada, consulte el Manual de redacción de patentes de la OMPI con ejemplos del manual a continuación, vea las páginas 84-90:

1. Composiciones de materia (ej. Un extracto, un compuesto).
2. Aparato (por ejemplo, un soporte para una cámara).
3. Métodos (por ejemplo, métodos para amplificar un ácido nucleico o para hacer té).
4. Proceso (por ejemplo, procesos para producir un producto en particular, como el té, conocido como Producto por Proceso).
5. Resultado a alcanzar / Parámetros. (por ejemplo, un cenicero que apaga automáticamente un cigarrillo).
6. Reclamaciones de diseño (por ejemplo, un diseño específico para un paraguas).

7. Patentes de plantas (limitadas a ciertas jurisdicciones, generalmente 1 reclamo para una variedad distinta de una variedad particular, por ejemplo, de **Banisteriopsis caapi** llamada 'Da Vine'). Restringido al cultivar reclamado y no debe confundirse con una patente de utilidad como en la controversia de la ayahuasca .
8. Las afirmaciones de biotecnología tienden a tomar la forma de "Un polinucleótido aislado seleccionado de ..." seguido de identificadores de secuencia (SEQ ID).
9. Utilice reclamaciones. En algunas jurisdicciones, un solicitante puede reclamar un nuevo uso para un compuesto conocido. Por ejemplo, el uso de un compuesto bien conocido para el tratamiento de una enfermedad (donde no se ha descrito previamente).
10. Reclamaciones de software. Las jurisdicciones también varían en cuanto a si permiten reclamaciones de software (y la ley también está sujeta a revisión). Los ejemplos incluyen referencias a "Un medio legible por computadora que almacena instrucciones ..." o "Una memoria para almacenar datos para el acceso de un programa de aplicación ..." seguido de más detalles sobre la estructura de datos y los objetos.
11. Afirmaciones ómnibus: tales como "1. Un aparato para cosechar maíz como se describe en la descripción. 2. Una máquina de jugo como se muestra en la Figura 4. "

Si bien esto suena como un montón de diferentes tipos de reclamaciones en la práctica, no encuentras todos estos todo el tiempo. En nuestra experiencia (principalmente trabajando en temas biológicos), las afirmaciones tienden a ser sobre composiciones de materia, incluida la biotecnología anterior, y métodos. Eso podría variar dependiendo de su campo de interés.



En el caso de la solicitud de patente de genomas sintéticos, podemos ver que estamos tratando con una reivindicación de método (reivindicación 1) relacionada con el ensamblaje de casetes de ácido nucleico. El primer reclamo es el más importante (y, a menudo, el más útil) de los reclamos porque todo lo que sigue normalmente depende de ese reclamo.

Las reclamaciones se pueden dividir en reclamaciones independientes y dependientes y forman un árbol de reclamaciones. En este caso, las reclamaciones 2 a 14 dependen de la reivindicación 1 y esto se puede identificar mediante la referencia a "El método de la reivindicación 1" al comienzo de cada una de estas

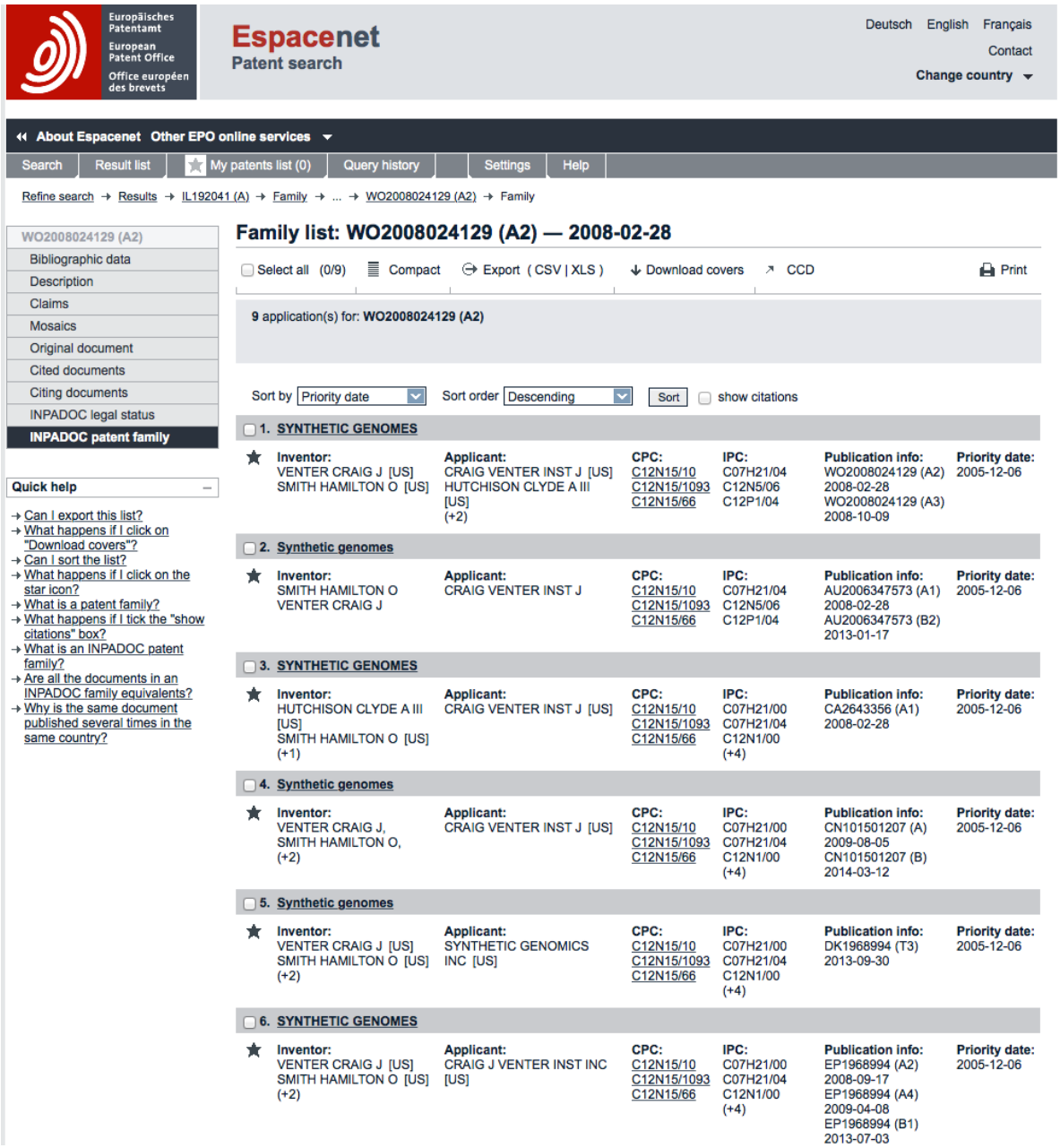
reivindicaciones. La siguiente reivindicación independiente en el documento WO2008024129A2aparece en la reivindicación 32 para "32. Un genoma sintético ". Los reclamos independientes no dependen de los otros reclamos.

Una nota final sobre las reclamaciones de patentes para el análisis de patentes es que las reclamaciones pueden cancelarse o modificarse en jurisdicciones particulares. En algunos casos, un examinador puede determinar que hay más de un invento en la solicitud. Esto puede hacer que la aplicación se divida en aplicaciones separadas vinculadas a la aplicación original (aunque las reglas varían en esto). Como tal, dependiendo del tipo de análisis requerido, puede ser importante rastrear a través de las aplicaciones en diferentes jurisdicciones. Aquí es donde entra la familia de patentes.

3.4.6 Miembros de la familia

En la discusión anterior, notamos que cuando se presenta una patente por primera vez en cualquier parte del mundo, se convierte en la presentación prioritaria o, como los autores tienden a llamar, la primera presentación. La presentación de prioridad también se convierte en el **padre** para cualquier seguimiento de las publicaciones en ese país u otro país (solicitudes y subvenciones, incluidas publicaciones administrativas como informes de búsqueda o documentos corregidos). La presentación prioritaria es, por lo tanto, el fundador de una **familia de patentes** y los documentos posteriores son niños que son **miembros de la familia**. Debido a que esto puede generar bastante confusión, veamos este ejemplo.

Miembros de la familia



Aquí podemos ver que la familia de nuestra solicitud internacional sobre genomas sintéticos contiene **9 aplicaciones** (incluido el documento de referencia de WO). Podemos ver que, en algunos casos, estas 9 solicitudes han llevado a más de una publicación a un total de **15 miembros de la familia**. Los números de publicación de patentes generalmente están acompañados por dos códigos de letras al final de los números llamados códigos de clase (por ejemplo, A2, B1, T1, etc.) Estos códigos en términos técnicos nos informan sobre el nivel de publicación pero también sobre el tipo de documento involucrado.

En algunos casos se trata de repúblicas administrativas. Por ejemplo, para el Código de tipo A3 del Tratado de Cooperación en materia de Patentes (WO) se entiende la publicación del informe de búsqueda internacional. Si bien este es un miembro de la familia, no nos gustaría incluir los recuentos de estos documentos en las estadísticas de patentes a menos que estemos estudiando las acciones de las oficinas de patentes.

En contraste, AU es el código de país de Australia y tiene dos documentos con códigos de clase A1 y B2. Debido a que las oficinas de patentes varían en el uso de estos códigos, pueden ser difíciles de interpretar con precisión, para obtener más detalles, consulte esta lista. En el caso del código amable de Australia, A nos dice que esta fue la publicación de una solicitud y que el código amable B nos dice que también se publicó como una concesión de patente. Cuando escaneamos la lista de miembros de la familia, podemos ver que hay otros países con publicaciones de tipo A y B.

La interpretación de los códigos de tipo requiere un cuidado considerable porque las prácticas de la oficina de patentes también varían con el tiempo. Por ejemplo, antes de 2001, la Oficina de Patentes y Marcas de los Estados Unidos solo publicaba documentos de patentes cuando se otorgaban y no publicaba solicitudes de patentes. Además, hasta 2001, la USPTO no usó códigos de clase o usó el código de clase A. A partir de 2001, la USPTO publicó ambas solicitudes y subvenciones con solicitudes que recibieron el código de clase A y otorga el código de clase B. El conocimiento de esto es fundamental para el cálculo de la patente. tendencias debido a que los datos anteriores a 2001 necesitan ser ajustados.

Dicho esto, **como regla general**, y con la excepción de los Estados Unidos antes de 2001, el código de clase A puede interpretarse como una solicitud y el código de clase B como una concesión de patente. Si bien enfatizamos que esto no es del todo satisfactorio, es el mejor proxy disponible para contar datos en todos los países hasta que las oficinas de patentes adopten prácticas más uniformes. Sin embargo, cuando se trata de un solo país, es mejor explorar la importancia de cada tipo de código.

Los datos de familia de patentes nos proporcionan una ruta para identificar todos los demás documentos que están vinculados a una primera presentación original. A través del entendimiento de las familias de patentes, también podemos avanzar en la distinción entre solicitudes de patentes y concesiones de patentes (aunque esto es imperfecto) con el propósito de desarrollar estadísticas. Si bien esto es satisfactorio para desarrollar el análisis de las tendencias de patentes, para otros fines, nos gustaría explorar otra información (por ejemplo, si se está manteniendo una concesión de patente ... ver más abajo).

Cuando se trabaja con datos de patentes, hay una variedad de tipos de familias de patentes. Por ejemplo, la base de datos de documentación de la EPO (DOCDB) es la fuente central de la mayoría de los datos de patentes y tiene un sistema de la familia DOCDB. Además, el Centro Internacional de Documentación de Patentes (INPADOC), ahora parte de la OEP, estableció el sistema INPADOC ampliamente utilizado. Las familias de la base de datos espacenet son un poco diferentes a las familias de INPADOC. Además, Thomson Reuters utiliza el sistema Thomson. Para una discusión en profundidad importante sobre las familias de patentes en relación con las estadísticas de patentes, consulte el excelente Documento de trabajo sobre ITS de la OCDE por Catalina Martinez (2010) Información sobre diferentes tipos de familias de patentes. Esto básicamente demuestra que las familias de DOCDB son un poco más pequeñas que las familias de INPADOC. Las familias de Thomson tienden a ser ignoradas en las estadísticas de patentes porque están limitadas a los usuarios comerciales de las plataformas de Thomson. Eso no significa que no deba usarlos, sino que si el desarrollo de un trabajo sobre tendencias de patentes que otros puedan seguir, entonces las familias DOCDB o INPADOC tienen mucho más sentido. En el trabajo del autor, tendemos a utilizar siempre los datos de la familia INPADOC cuando están disponibles.

Por ahora, esto puede sonar bastante complicado. En la práctica, no lo es. Una forma muy simple de entender una familia de patentes es la siguiente.

Una familia de patentes es una pila de documentos con el padre (prioridad) en la parte inferior de la pila. Esos documentos pueden haber sido publicados en varios países y en diferentes idiomas, pero como se vinculan con el mismo padre (prioridad) son miembros de su familia.

Este enfoque simple para comprender una familia de patentes también es muy útil cuando se piensa en qué contar.

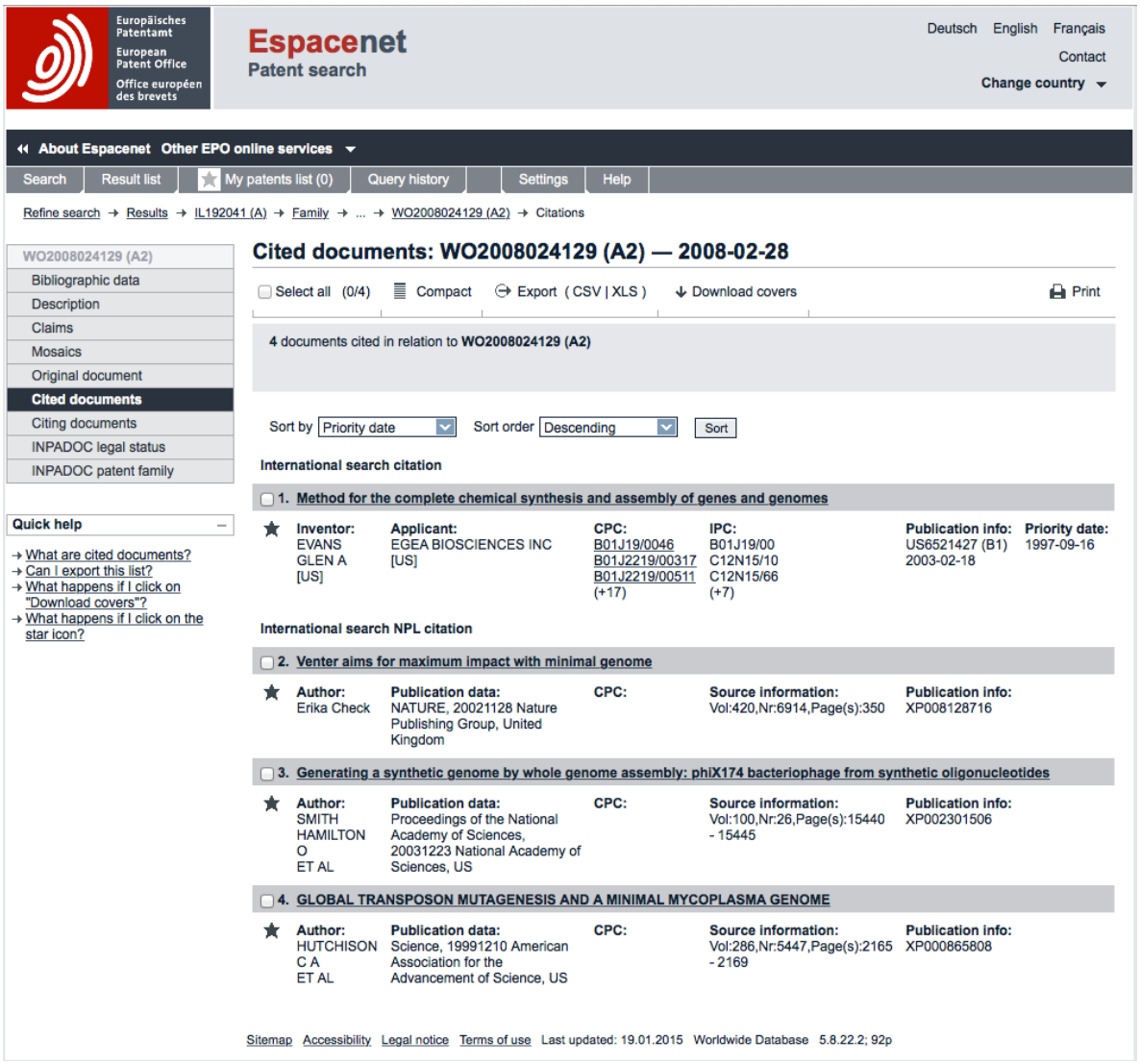
1. Cuando desarrollamos conteos basados en **familias de patentes,** contamos las primeras solicitudes de una solicitud de patente y nada más. Es decir, el documento en la parte inferior de cada pila.
2. Cuando contamos a **los miembros de la familia de patentes,** estamos contando todos los documentos que enlazan a la familia de patentes como su padre. Es decir, todos los documentos en la pila.

El enfoque anterior le permitirá contar con éxito miles o millones de documentos de patentes de una manera que tenga sentido para usted y para otros. Sin embargo, tenga en cuenta que en algunos casos una patente puede tener más de un padre prioritario. Esto parece ser particularmente cierto para las patentes de software. Es decir, nos enfrentamos a una relación de "muchos a muchos" en lugar de a una

relación de "uno a muchos" más simple entre los documentos de prioridad y los miembros de la familia. Por lo tanto, se necesitarían medidas adicionales para desarrollar conteos para abordar este aspecto de los datos.

3.4.7 Citado

Se puede decir que los solicitantes de patentes están "de pie sobre los hombros de gigantes" para tomar prestado del trabajo de la economista Suzanne Scotchmer sobre patentes. Para nuestros propósitos, **los juicios sobre la novedad y el paso inventivo durante el examen se basan en evaluaciones de la bibliografía de patentes existente (lo que otros han afirmado anteriormente) y lo que se denomina Literatura no patentada (NPL), incluidas publicaciones científicas y otros materiales que constituyen Art**". La existencia del estado de la técnica puede significar que una solicitud de patente no puede continuar o que los solicitantes deberán limitar lo que afirman a los aspectos de la invención que no existen en el estado de la técnica. Esta información se registra en el campo Documentos citados en bases de datos como espacenet.



En este caso, los documentos citados incluyen una subvención US6521427B1 de EE. UU. De 2003 a Egea Biosciences con una presentación prioritaria en 1997 relacionada con la síntesis de oligonucleótidos para "el ensamblaje de genes y genomas de organismos artificiales completamente sintéticos" utilizando la síntesis de genes dirigida por computadora. Además, los documentos citados incluyen literatura citada (otorgada un código XP en espacenet) que en dos casos se originan en los inventores de la aplicación de genomas sintéticos.

La bibliografía de patentes citadas y no patentes puede tener dos fuentes.

1. Información proporcionada por los solicitantes.
2. Documentos identificados por los examinadores durante la búsqueda y / o examen. En algunos países, los solicitantes deben proporcionar información detallada sobre el estado de la técnica relevante para la invención

> reivindicada. En otros casos el requisito es más débil. Como podríamos esperar, los solicitantes se mostrarán reacios a divulgar información que invalide o complique enormemente sus esfuerzos para obtener una patente. Además, los examinadores también realizarán búsquedas para identificar el arte relevante, pero los requisitos de los examinadores para divulgar realmente esa información también pueden variar. Para una discusión, vea el trabajo de Colin Webb y sus colegas de la OCDE aquí y el Manual de estadísticas de patentes de la OCDE de 2009, en particular, el Capítulo 6. Las citas agregadas por los examinadores son generalmente más importantes que las agregadas por los solicitantes y en algunos casos pueden estar marcadas en las bases de datos de patentes. En este caso particular, podríamos encontrar información adicional sobre el origen de las citas consultando el documento original para la publicación del informe de búsqueda internacional (documento A3) mencionado anteriormente en el documento WO2008024129A3 . Como podemos ver, esto contiene la página principal y luego un conjunto de citas acompañadas por una categoría en la que se considera que la entrada marcada con X para la concesión de la patente a Egea Biosciences afecta las afirmaciones de novedad y / o actividad inventiva cuando se toma por sí sola.

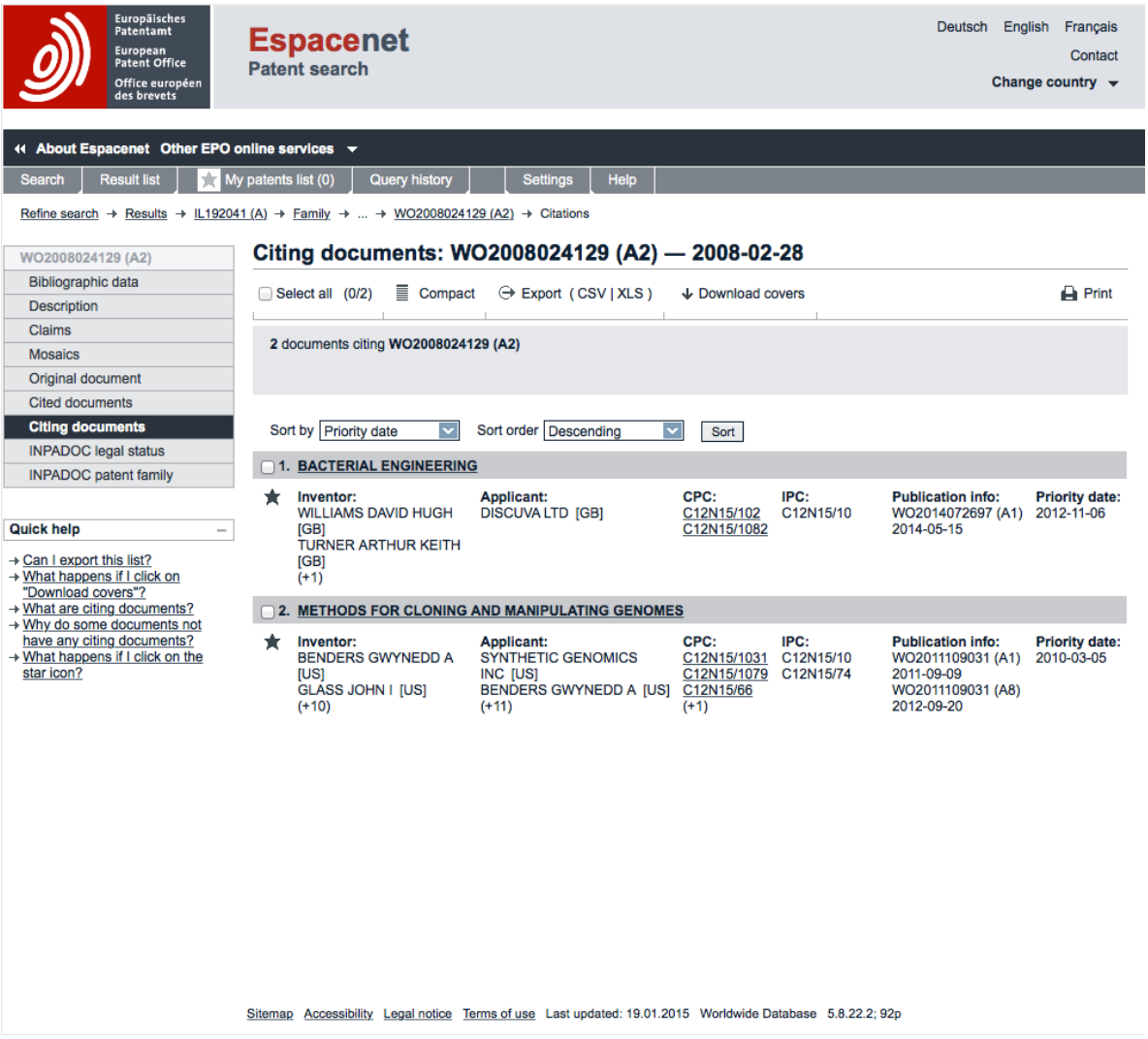
En la práctica, los solicitantes pueden usar citas para ajustar sus reclamos y, como tales, no son necesariamente un obstáculo para obtener una concesión de patente (como hemos visto en los datos de la familia de patentes). Sin embargo, dependiendo de nuestro propósito, los datos de citas son muy útiles en el análisis de patentes.

1. Permite recopilar datos relevantes de patentes que podrían haberse perdido debido a las limitaciones de una consulta de búsqueda en particular. Por lo tanto, ayuda a completar la imagen para un análisis de panorama de patentes o búsquedas de la técnica anterior relevante.
2. Para la investigación académica, puede mostrar la actividad que influyó en la aparición de un campo en particular, como la biología sintética.

Al revisar la bibliografía de patentes citadas y no patentes, tenga en cuenta que un documento de patente citado (que puede ser una solicitud o una subvención) no puede caer directamente en el campo de la invención de interés. Por ejemplo, una característica particular de una invención reivindicada en un campo de tecnología (como la óptica militar) puede afectar los desarrollos en otro campo (como la óptica médica) o aparentemente no tener ninguna relación, excepto por un aspecto técnico específico.

3.4.8 citando

Citar datos es lo opuesto a los datos citados. Una manera útil de pensar esto es que los datos citados significan citas anteriores mientras que los datos citados significan citas futuras. Los datos de citación o las citas posteriores son solicitudes de patente posteriores que citan nuestro documento de referencia de la siguiente manera.



En este caso, hay dos documentos citando en el momento de la escritura. Uno es del solicitante británico Discuva para ingeniería bacteriana WO2014072697A1 . Un segundo es de Synthetic Genomics (un brazo comercial del J. Craig Venter Institute) para métodos de clonación y manipulación de genomas con algunos de los mismos inventores enumerados en la solicitud WO2011109031A1 .

Las citas a plazo proporcionan información sobre los solicitantes que están siendo afectados por una solicitud o concesión de patente en particular o, en mayor escala, por conjuntos de documentos. Estos datos se pueden utilizar de manera estratégica para identificar a otras personas que trabajan en un campo en particular que está "cerca" del área de interés de una empresa o universidad. Esta información podría informar decisiones sobre la creación de alianzas potenciales o, en otras circunstancias, podría informar decisiones sobre procedimientos de infracción.

En términos más amplios, los datos de citación a futuro pueden informar el análisis del panorama de patentes sobre el desarrollo de un campo en particular (como la biología sintética), mientras que teniendo en cuenta que la actividad de patentes en un campo puede tener efectos secundarios en otras áreas de la tecnología aparentemente no relacionadas.

3.4.9 Estado legal

Como se mencionó anteriormente, los códigos de tipo de patente al final de los números de publicación proporcionan una indicación del nivel de publicación y el tipo de documento de patente. En los casos en que esto involucra códigos de clase específicos (por ejemplo, B), a menudo es un indicador de una concesión de patente. Sin embargo, para obtener información adicional, necesitamos revisar los datos del estado legal como se muestra a continuación.

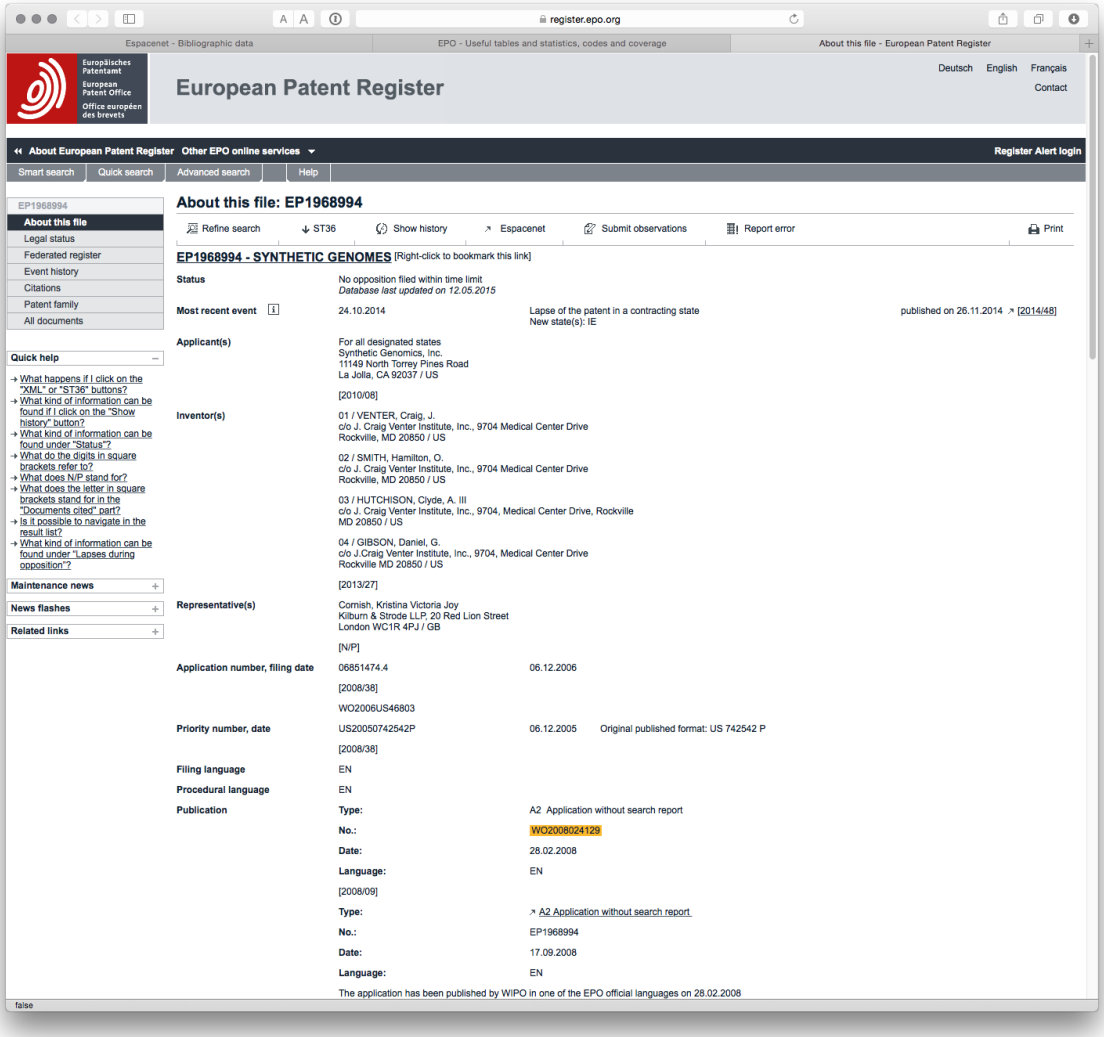
En este caso, el punto más obvio acerca de los datos es que nos informa que la aplicación está entrando en la fase nacional en varios países diferentes. Es decir, los solicitantes están siguiendo la solicitud en los países específicos que figuran en la página de inicio (arriba) en el campo de Estados designados (todos los Estados contratantes del PCT se enumeran de forma predeterminada). Como tales, los solicitantes están señalando su intención de buscar subvenciones de patentes en estos países. En otro caso, los datos de estado legal pueden indicar que una solicitud ha sido rechazada, que una patente otorgada ha caducado debido a la falta

de pago de las tarifas o ha caducado. Se puede obtener información adicional a través de la interpretación de los códigos de estado legal con más información disponible descargando la Categorización de los códigos de estado legal recientemente usados del sitio web de la OEP.

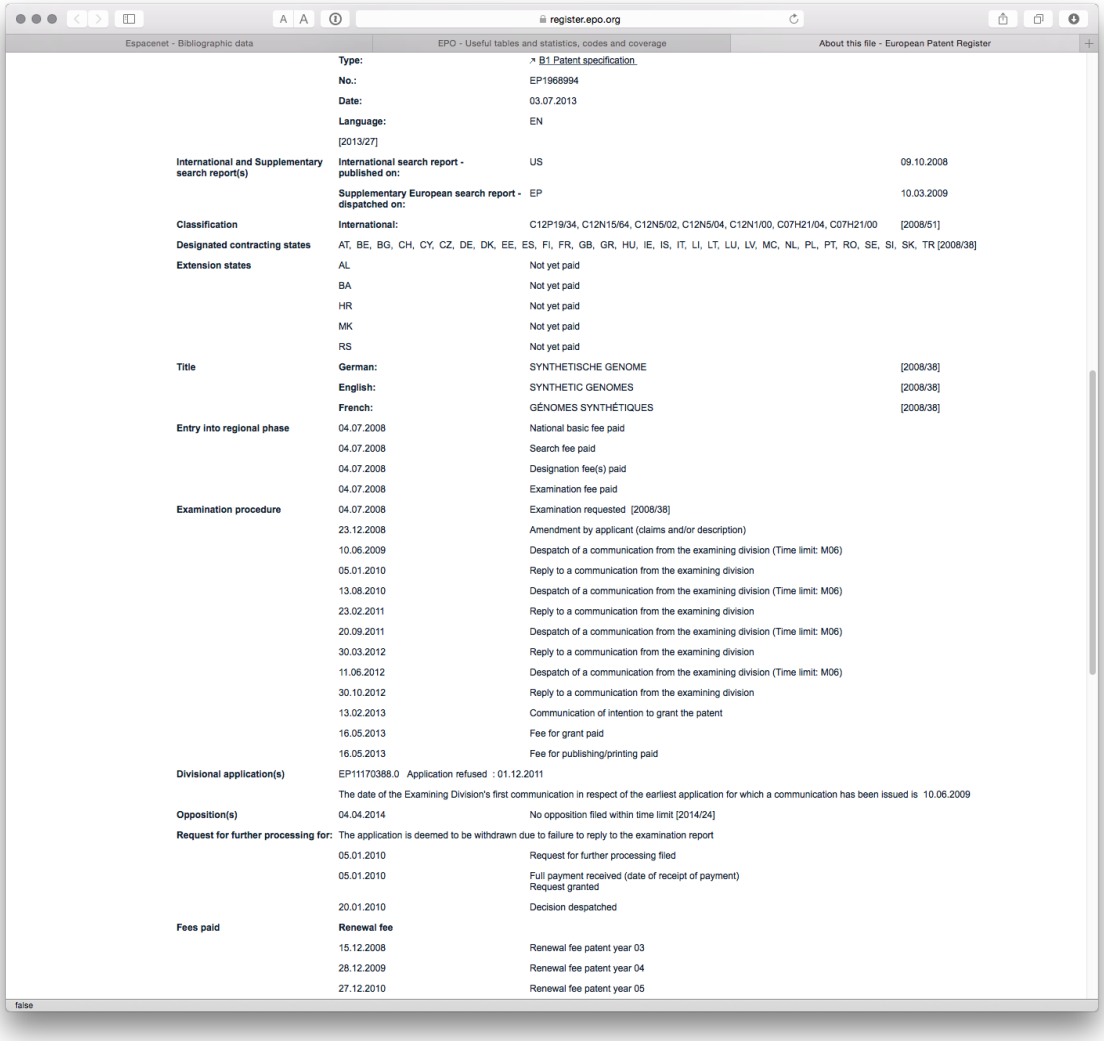
Al revisar los datos del estado legal, tenga en cuenta que puede no ser reciente o completa. Por esta razón, la investigación a nivel nacional (y la consulta con un profesional de patentes) generalmente será necesaria para determinar lo que está sucediendo con una solicitud o subvención en particular.

Registros de Patentes

La información adicional sobre un documento de patente suele estar disponible consultando registros de patentes a nivel nacional o regional. En el caso de las aplicaciones a nivel europeo, normalmente hay más información disponible a través del botón Registro EP en la página principal. Si seleccionamos esto para nuestro documento de WO, seremos trasladados a la entrada de registro de EP para el miembro de la familia europea EP1968994 . Como a continuación.



A partir de esta información podemos ver que no se presentó ninguna oposición a la solicitud de patente dentro del límite de tiempo. Luego vemos que el evento más reciente es un lapso de una patente en Irlanda (IE) junto con el historial de publicaciones. Si nos desplazamos hacia abajo la página más información queda disponible.



En este caso, obtenemos acceso a la información sobre el historial de comunicaciones escritas entre los solicitantes y la OEP junto con detalles del pago de las tarifas de renovación de patentes. Fuera de la vista en esta imagen se encuentran los datos de citas, incluidos los DOI de la literatura citada que se vincularán directamente con el artículo en cuestión cuando esté disponible.

En el menú a la izquierda hay información adicional disponible. El menú de registro federado proporciona acceso a los registros nacionales de patentes de los estados contratantes designados en virtud del Convenio Europeo de Patentes, como puede verse aquí .

Finalmente, el elemento del menú Todos los documentos proporciona acceso a las copias de la correspondencia disponible y otros documentos que se pueden

descargar como un archivo Zip. También es posible enviar una observación de un tercero usando el botón enviar observaciones en el menú.

Los datos dentro del registro pueden ser particularmente útiles para explorar el historial y el estado de una aplicación, como la modificación de las reivindicaciones de patentes a la luz de los informes de búsqueda. También es muy útil para identificar y revisar la oposición a una aplicación en particular.

Dentro de Europa es muy fácil consultar los detalles del registro. Para ayudar a acceder a la información de registro en otros países, la OMPI lanzó recientemente un portal de registro de patentes para simplificar la tarea de ubicar el registro de patentes en los países de interés.

3.5 Redondear

En este capítulo, hemos recorrido algunos de los campos de datos de patentes más importantes utilizando un solo ejemplo y la base de datos espacenet. Como ahora se puede apreciar, una comprensión básica de los campos de datos de patentes abre una gran cantidad de información adicional sobre un único documento de interés.

Estos campos básicos son también los componentes básicos para el análisis sofisticado de patentes. En futuros capítulos nos centraremos en:

- Recuperando datos con estos campos
- Limpiando los datos en estos campos.
- Tendencias de mapeo
- Mapeo de red
- Mapeo geográfico

Capítulo 4 Conjuntos de datos

En este capítulo presentamos los conjuntos de datos de patentes desarrollados para el Proyecto Open Source Patent Analytics como conjuntos de capacitación para el análisis de patentes. Los conjuntos de datos se utilizarán en los tutoriales. Los conjuntos de datos crecerán con el tiempo, pero los presentaremos brevemente y explicaremos cómo acceder a ellos.

Los conjuntos de datos se encuentran en el repositorio de GitHub del proyecto . Para descargar archivos individuales, haga clic en el enlace y luego seleccione sin formato para descargar el archivo.

4.1 Los conjuntos de datos

Los conjuntos de datos tienen la intención de ilustrar el rango de posibilidades para los datos de patentes, incluidos algunos de los desafíos que pueden surgir al limpiar y analizar los datos de patentes. También se extraen de diferentes fuentes.

4.1.1 Conjuntos de datos de patentes de pizza

A casi todos les gusta la pizza y es fácil buscar en la base de datos de patentes el término "pizza". También es un área de actividad de patentes que abarca una amplia gama de tecnologías como hornos de pizza, cajas de pizza, cortadores de pizza y aderezos de pizza, etc. Por lo tanto, es útil para demostrar formas de interrogar datos de patentes para temas particulares.

1. pizza_smalles un conjunto de datos muy pequeño de 26 filas creado al descargar la primera página de resultados de la base de datos espacenet de la Oficina Europea de Patentes para una búsqueda inteligente en "pizza". Es un conjunto de datos de prueba rápido y fácil.
2. pizza_medium se creó a partir de una muestra de datos de una búsqueda en la base de datos de WIPO Patentscope para el término "pizza" y contiene 9,996 filas de datos. Está pensado para ilustrar el formato de datos de Patentscope y para permitir el trabajo en un conjunto de datos de tamaño mediano. Tenga en cuenta que el formato varía del formato espacenet y presenta diferentes desafíos. Una característica importante de los datos de Patentscope desde un punto de vista estadístico es que el campo marcado publication_numberen los datos originales carece de un código de tipo de dos letras y, por lo tanto, es un application_number.
3. El pizza_medium_cleanconjunto de datos es una versión predefinida del pizza_mediumconjunto de datos. Específicamente, el campo de solicitantes

e inventores ya se ha limpiado junto con los caracteres corruptos y otras tareas de limpieza comunes. Esto facilita el trabajo con los datos y este conjunto de datos es el conjunto de datos central del Manual. Como se mencionó anteriormente, tenga en cuenta que el campo Número de publicación de Patentscope se refiere más adecuadamente a un número de solicitud en ausencia de un código amable.

4. pizza_slicedes un conjunto de cinco archivos **.csv** para una búsqueda de pizza en espacenet . Está diseñado para ilustrar los problemas relacionados con la carga de múltiples archivos en R. También ilustra problemas con la corrupción de caracteres y la importancia de la limpieza previa de los datos antes del análisis.
5. pizza_lens_1000es un conjunto de datos en bruto de 1000 registros que incluye el término pizza descargado de la base de datos de The Lens. El conjunto de datos no se ha limpiado.

4.1.2 Conjuntos de datos de Patentes del paisaje

Tres conjuntos de datos se extraen de los Informes de Patentes de la OMPI . Los conjuntos de datos abordan diferentes temas, presentan una variedad de campos y formatos y son de diferentes tamaños. Cada conjunto de datos está vinculado a un informe detallado del panorama de patentes que proporciona una perspectiva de los enfoques para el análisis de patentes.

1. Ewaste presenta los resultados de la investigación para un informe sobre la actividad de patentes para el reciclaje de desechos electrónicos y sus implicaciones para los países en desarrollo.
2. solar_cooking presenta los datos que respaldan un informe sobre tecnologías que utilizan la energía solar como fuente para cocinar y pasteurizar alimentos.
3. Ritonavir presenta los datos de un informe de patente sobre la actividad de patentes para el medicamento antirretroviral para el VIH Ritonavir en el campo de los productos farmacéuticos. El conjunto de datos ilustra la actividad específica en torno a cuestiones como la dosificación y también el problema de la "perennidad" en la actividad de patentes.

4.1.3 Otros conjuntos de datos

1. wipo es una única hoja de datos de Excel sobre tendencias en solicitudes de patentes y tasas de crecimiento de los Indicadores de propiedad intelectual mundiales de la OMPI - Edición 2014 . Los datos se utilizan para la representación gráfica simple en herramientas como R e ilustran la necesidad de omitir filas al leer datos en herramientas de análisis.

2. WIPO_sequence_data. Este conjunto de datos contiene una pequeña muestra de los datos de secuencia del año 2000 disponibles de forma gratuita en la base de datos de la OMPI Patentscope . Este conjunto de datos se puede utilizar para explorar el análisis de datos de secuencias de patentes.
3. Biología sintética . Esta es una muestra de datos de Thomson Innovation desarrollados por Paul Oldham para la investigación de la actividad de patentes relacionadas con la biología sintética. Los datos se han limpiado exhaustivamente en VantagePoint de Search Technology Inc. y se pretende ilustrar el uso de los datos de una base de datos de patentes comerciales.

4.1.4 Redondear

La sección de conjuntos de datos del proyecto proporciona una serie de conjuntos de capacitación útiles de una variedad de fuentes y que muestra una variedad de características. Estos son conjuntos de datos de acceso abierto que se pueden usar para probar diferentes enfoques, pero por favor acredite sus fuentes. Es posible que se agreguen más conjuntos de datos a la versión en línea del Manual a su debido tiempo. Estamos particularmente interesados en los datos de muestra de STN, QuestelOrbit, PATSTAT u otros proveedores de datos que se pueden usar como conjuntos de capacitación.

Capítulo 5 Bases de datos

5.1 Introducción

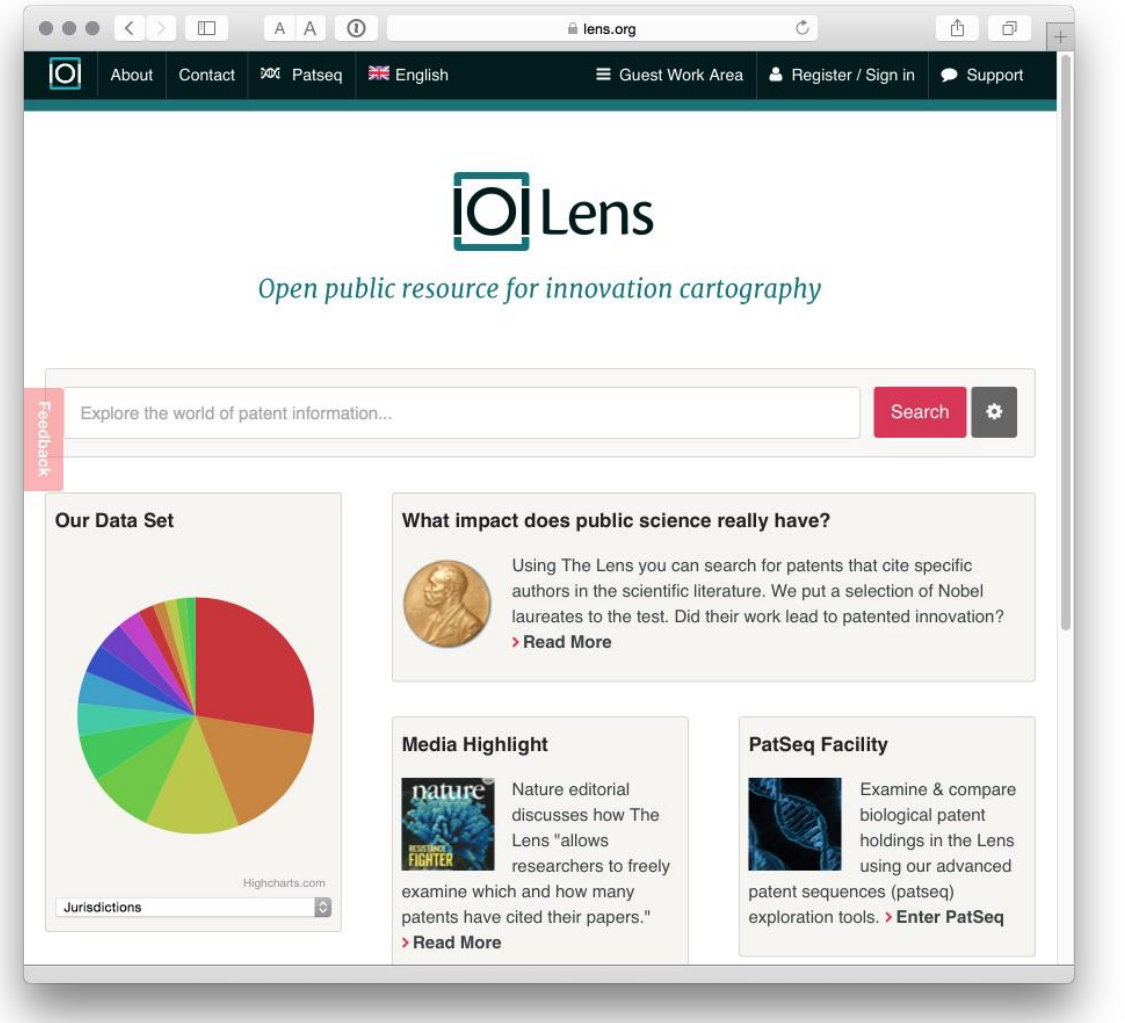
Este capítulo proporciona una descripción general rápida de algunas de las principales fuentes de datos de patentes gratuitos. Está diseñado para una referencia rápida y señala algunas herramientas gratuitas para acceder a bases de datos de patentes con las que puede que no esté familiarizado.

No hace falta decir que obtener acceso a los datos de patentes en primer lugar es fundamental para el análisis de patentes. Hay bastantes servicios gratuitos y destacaremos algunos de los más importantes. La mayoría de las fuentes libres tienen fortalezas o debilidades particulares, como el número de registros que se pueden descargar, los campos de datos que se pueden consultar, el formato en el que se devuelven los datos o cómo cleanse encuentran los datos en términos de las horas requeridas para prepararse para el análisis. No entraremos en todos los detalles, pero proporcionaremos algunos consejos básicos.

5.2 Las bases de datos

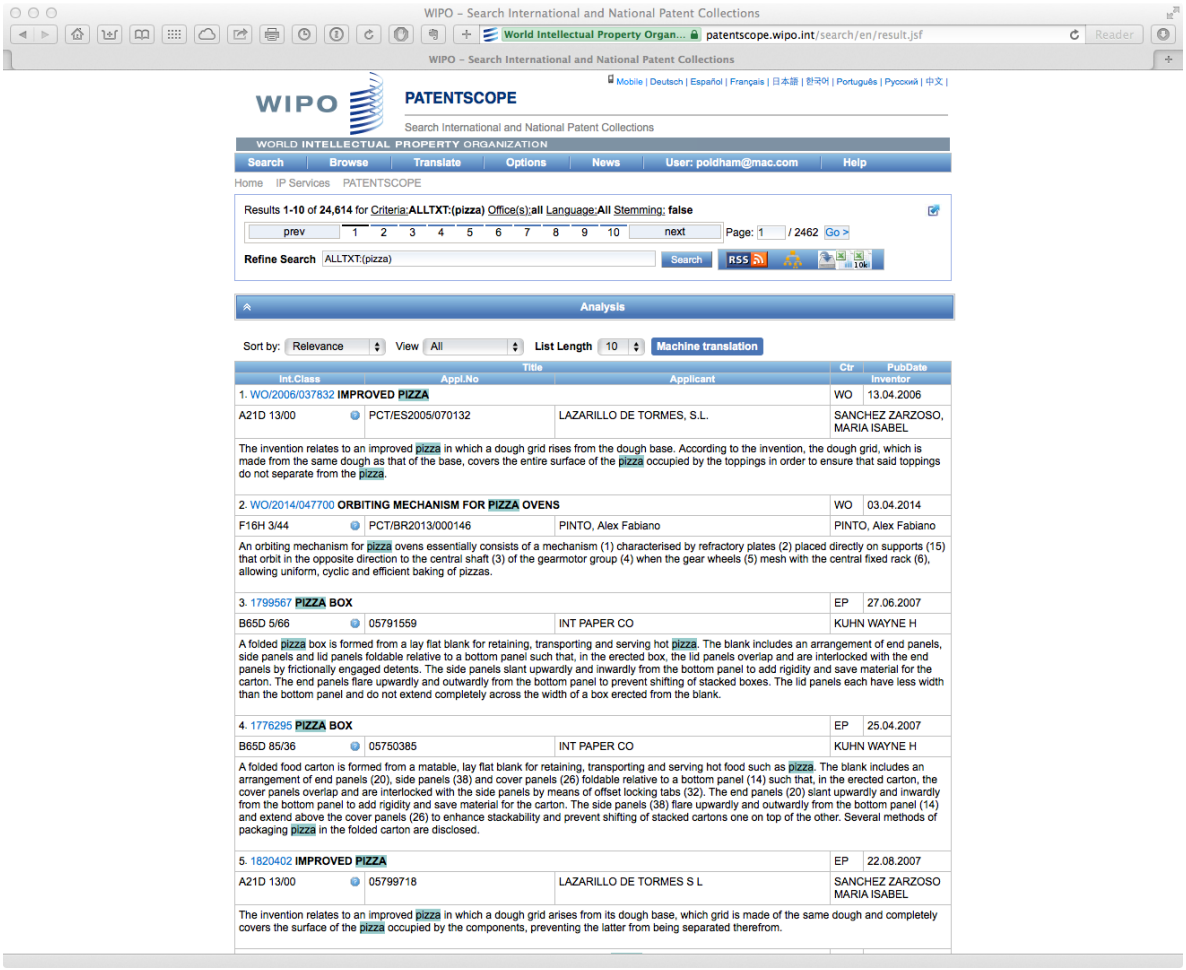
5.2.1 Lens

Anteriormente conocido como Lens Patent, este es un sitio bien diseñado con bastantes opciones de visualización y acceso a datos de secuencia. Es posible buscar el título, el resumen, la descripción y las reclamaciones de los documentos de patentes y crear y compartir datos en colecciones. En 2015 se agregó la capacidad de descargar hasta 10,000 registros a la vez. Cuando se combina con gráficos interactivos que permiten al usuario profundizar en el conjunto de resultados, esto ha transformado el objetivo en una base de datos y una herramienta de visualización muy útil e innovadora.

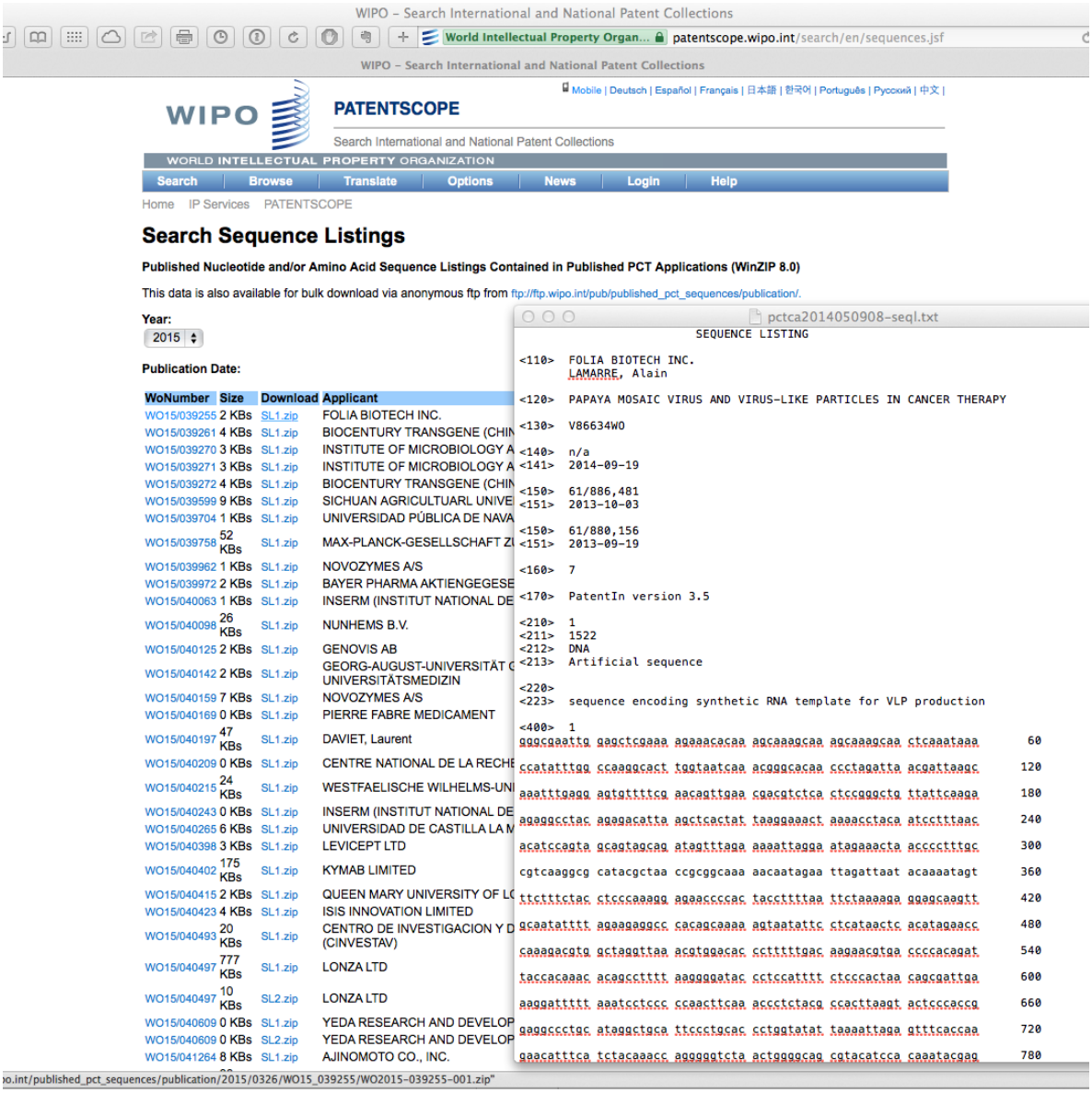


5.2.2 Patentscope

La base de datos de la OMPI Patentscope proporciona acceso a los datos del Tratado de Cooperación en materia de Patentes, que incluyen descargas de una selección de campos (hasta 10.000 registros), una herramienta de traducción de expansión de búsquedamuy útil y traducción .

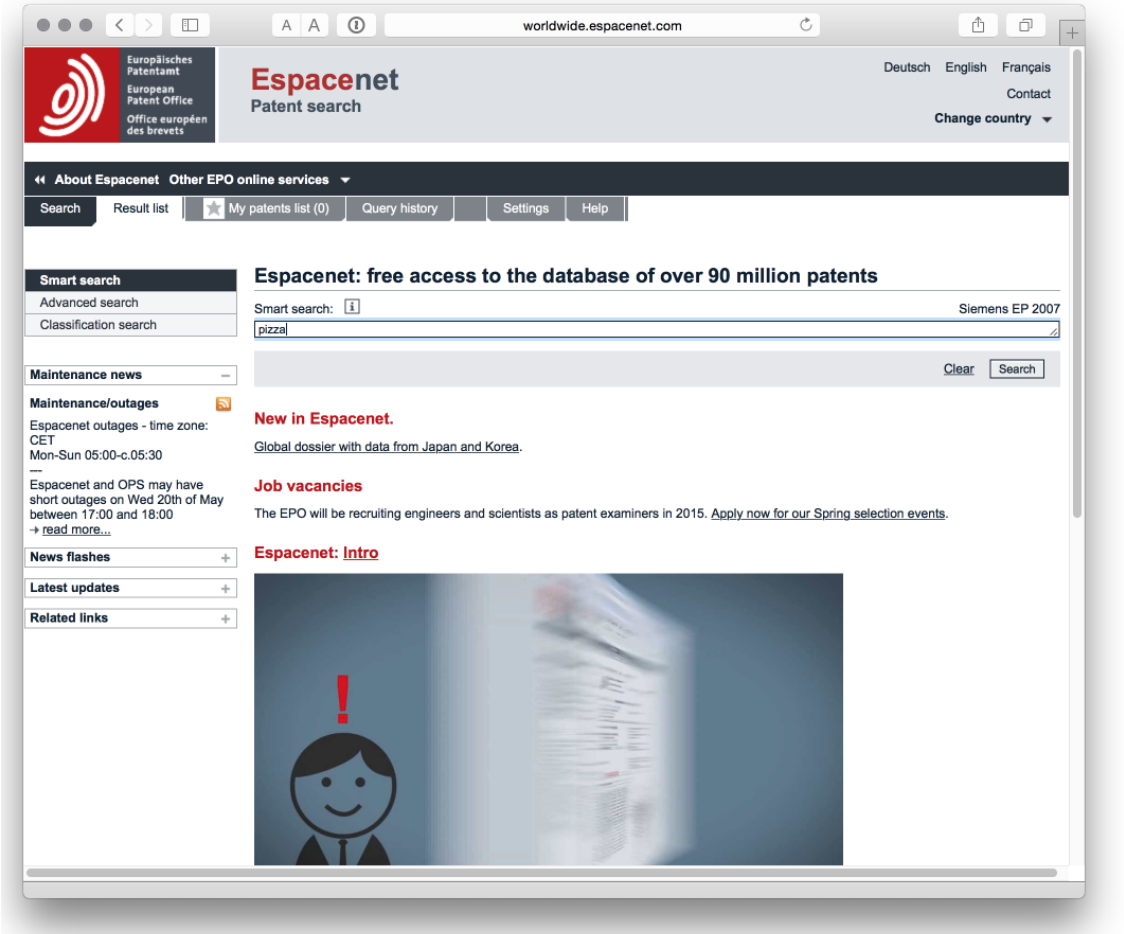


Obtención de datos de secuencia de Patentscope. Tenga en cuenta que esto se convierte rápidamente en gigabytes de datos.



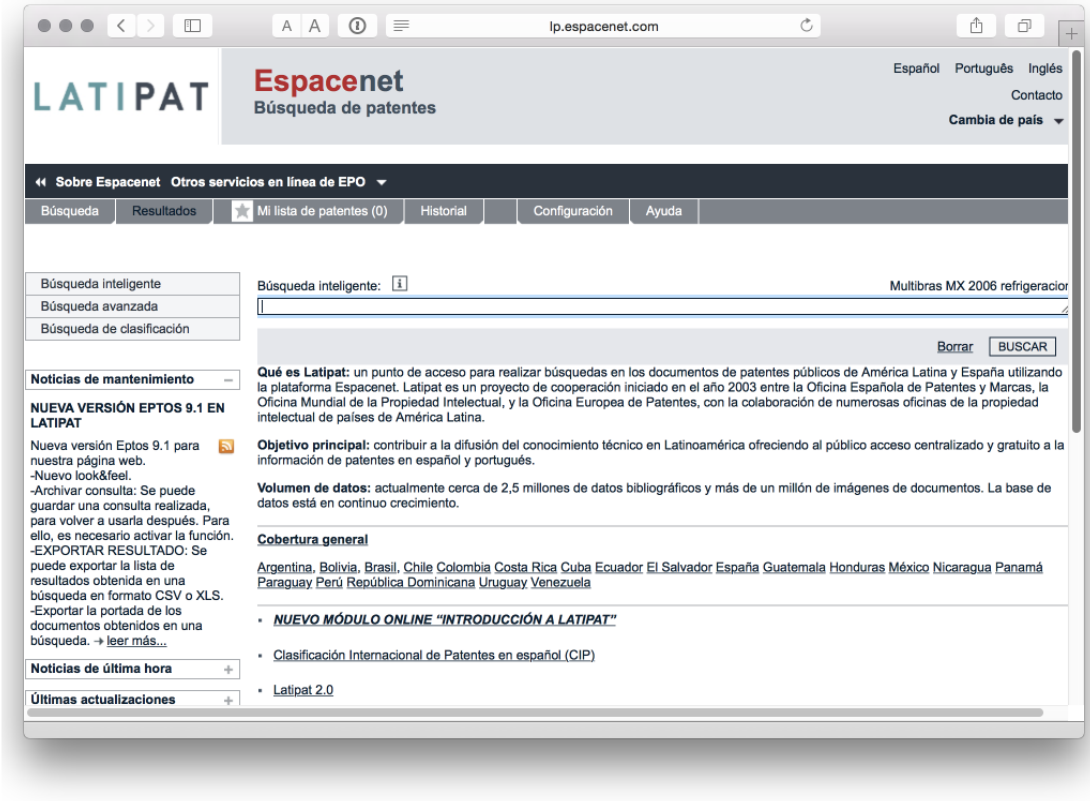
5.2.3 espacenet

Probablemente la base de datos de patentes gratuita más conocida de la Oficina Europea de Patentes.



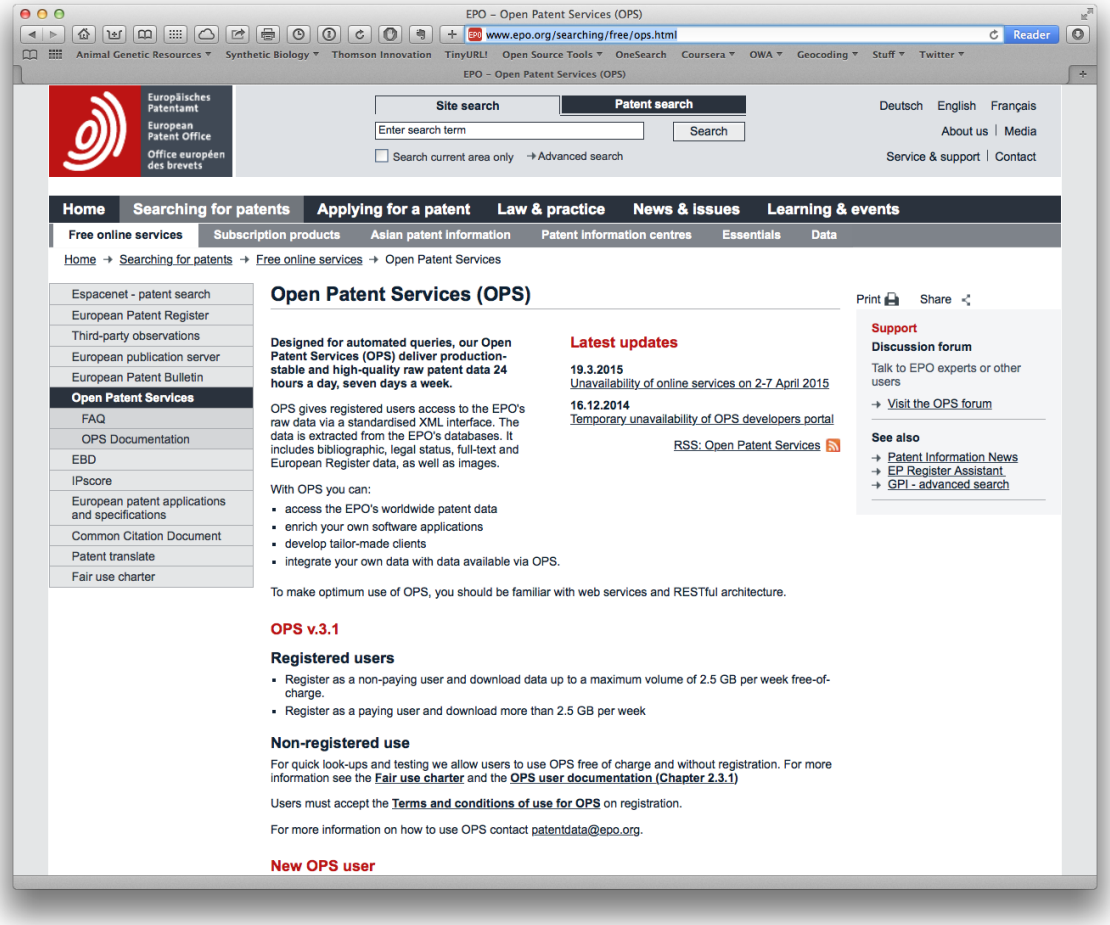
5.2.4 LATIPAT

Para los lectores de América Latina (o España y Portugal), LATIPAT es un recurso muy útil.

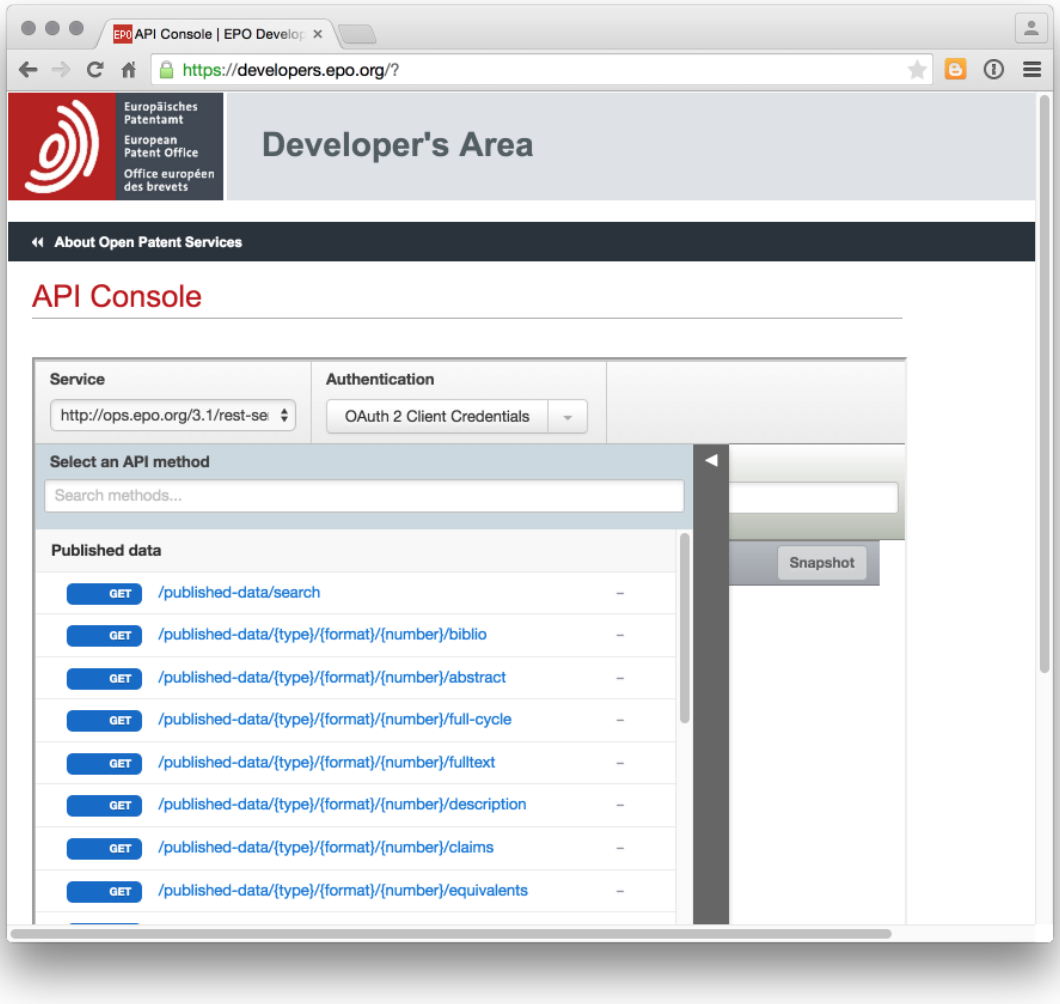


5.2.5 Servicios de patentes abiertas de la OEP

Acceda a los datos de patentes a través de la Interfaz de programación de aplicaciones (API) de EPO de forma gratuita. Requiere conocimientos de programación.



El portal de desarrolladores le permite probar sus consultas de API y se recomienda.

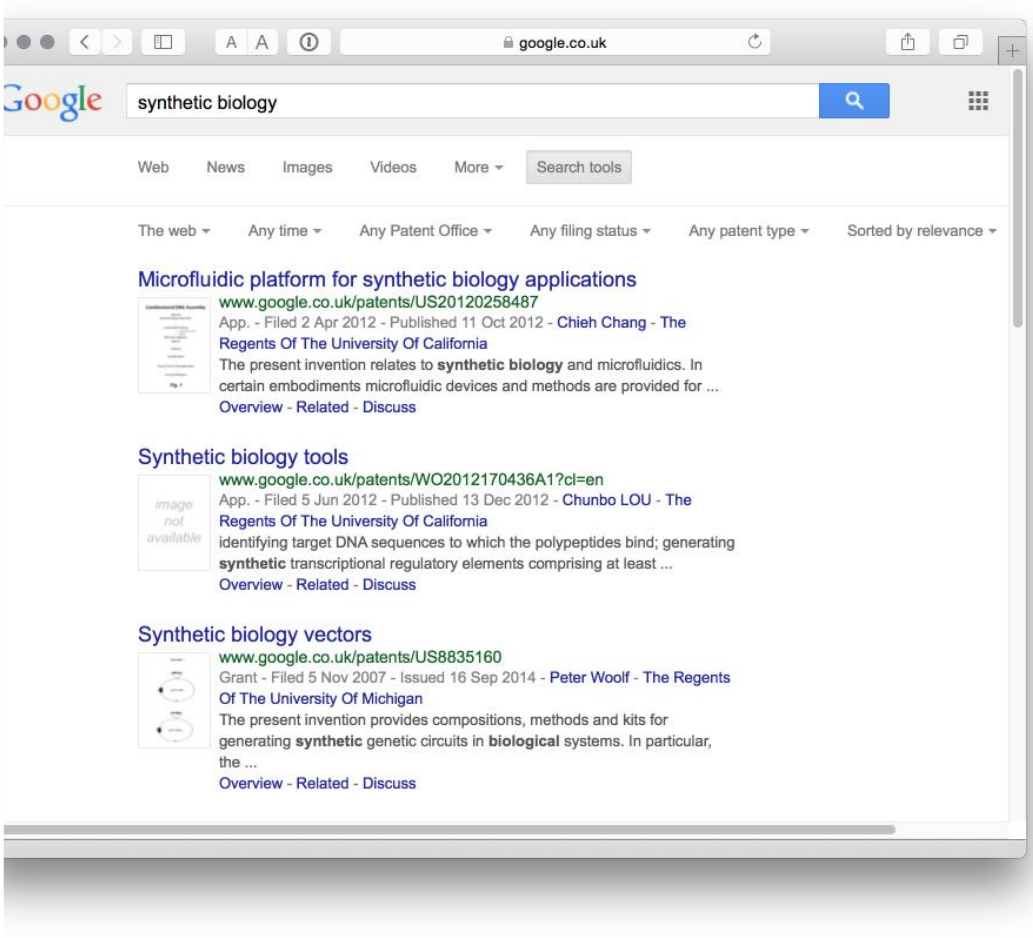


5.2.6 Vista de patentes de la USPTO

La página de búsqueda de la base de datos principal de la USPTO también se puede describir razonablemente como ... antigua. En 2016, el equipo de la USPTO inició una iniciativa de datos abiertos y movilidad que abre los datos de patentes y marcas de la USPTO. El nuevo Portal de fecha abierta todavía está en Beta, pero proporciona una visión de las cosas por venir.

Como parte del cambio para abrir datos, la USPTO ha establecido una Vista de Patentes externa para búsquedas gratuitas y descargas masivas . Si la búsqueda simple no satisface sus necesidades, o las opciones masivas son demasiado abrumadoras, es probable que el nuevo servicio JSON API satisfaga sus necesidades. Los servicios aún están en fase beta, pero este es un desarrollo muy interesante para aquellos que necesitan mayores niveles de acceso a datos de patentes o acceso a campos de datos específicos.

5.2.7 patentes de Google

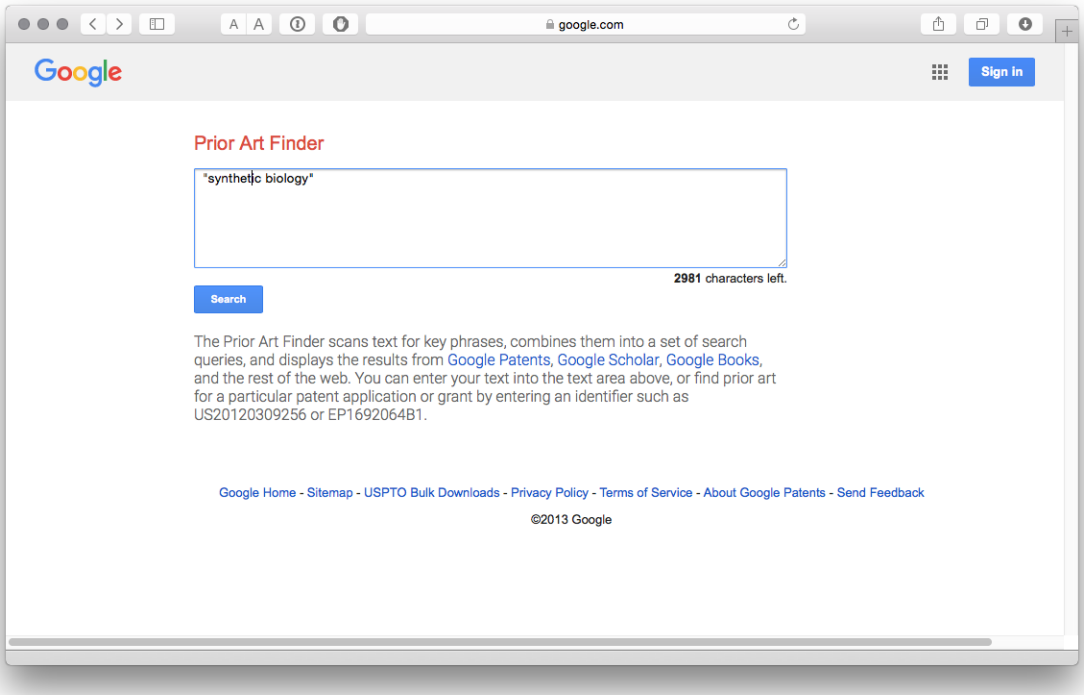


La API de búsqueda de patentes de Google ha quedado en desuso. El acceso a través de la API de Google Custom Search API con la bandera de patentes informó que &tbm=ptscon el ejemplo de código para el uso de la API de Python.

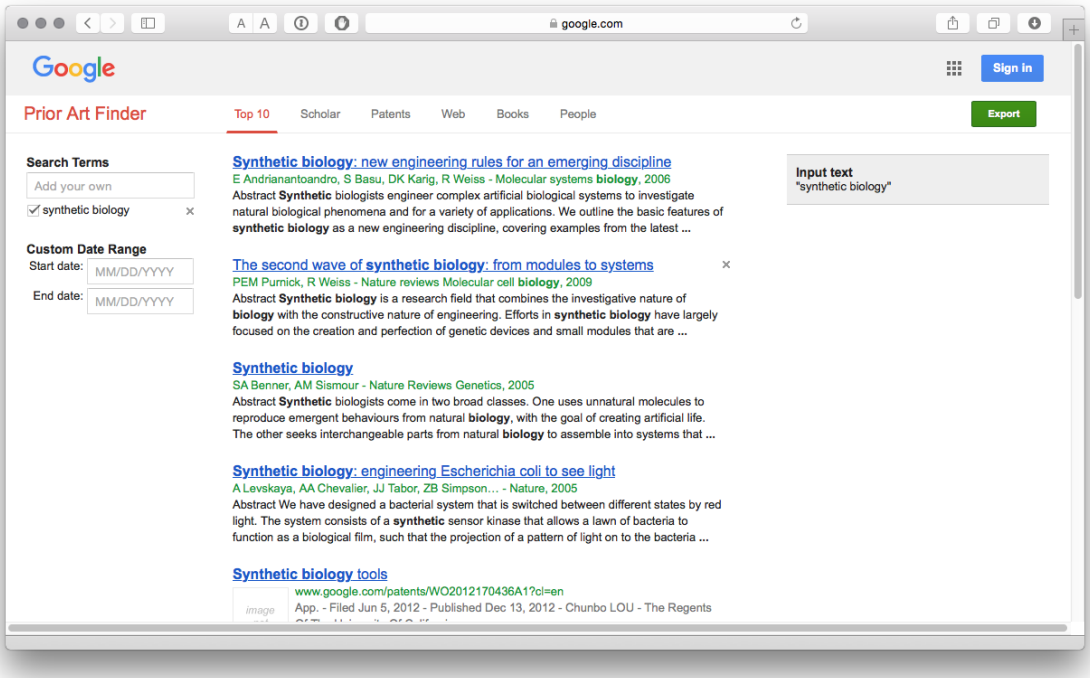
En la versión gratuita de Google Custom Search API, la recuperación de datos es limitada y los encabezados de los campos de patentes no están claros (es decir, utilizan nombres no estándar). Para el análisis de patentes gratuito, la Búsqueda personalizada de Google actualmente tiene un uso muy limitado.

5.2.8 Buscador de arte previo de Google

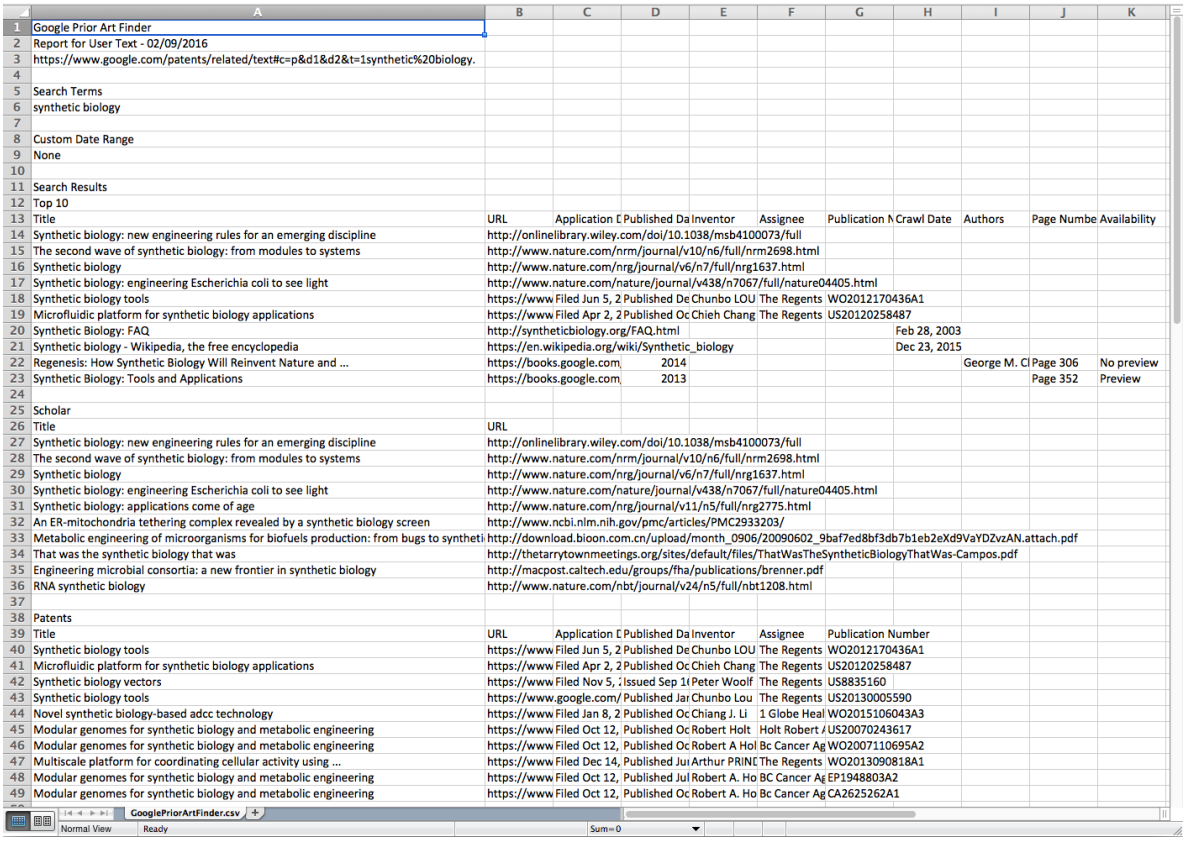
El Buscador de Arte Anterior de Google es un desarrollo relativamente reciente que le permite ingresar términos de búsqueda o números de patentes y ver y exportar resultados.



Los resultados incluyen un Top Ten y se desglosan en secciones que incluyen Google Scholar, patentes, etc.



El botón Exportar exportará los diez resultados principales para cada sección en un archivo .csv.

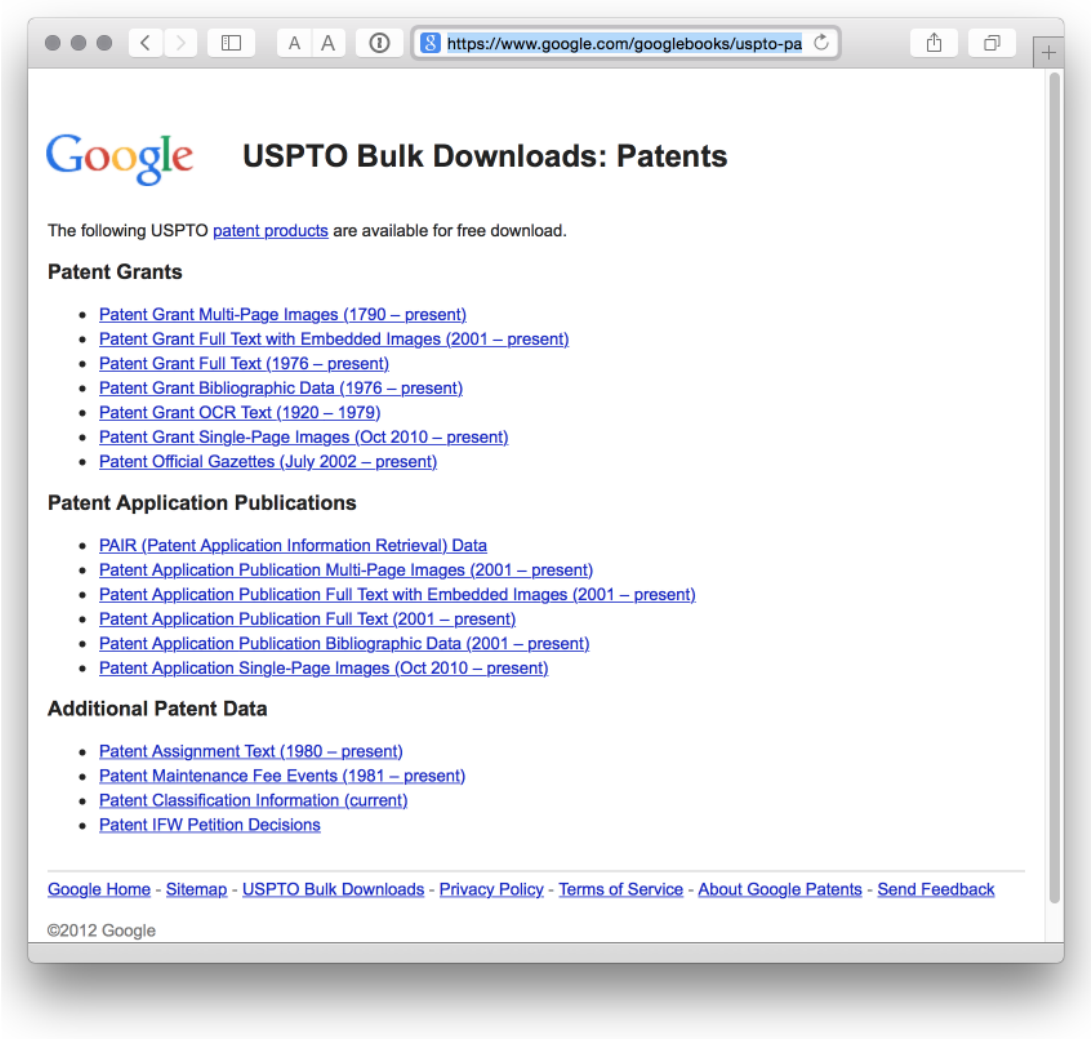


	A	B	C	D	E	F	G	H	I	J	K
1	Google Prior Art Finder										
2	Report for User Text - 02/09/2016										
3	https://www.google.com/patents/related/text#c=p&d1&d2&t=1synthetic%20biology.										
4											
5	Search Terms										
6	synthetic biology										
7											
8	Custom Date Range										
9	None										
10											
11	Search Results										
12	Top 10										
13	Title	URL	Application D	Published Da	Inventor	Assignee	Publication N	Crawl Date	Authors	Page Numbe	Availability
14	Synthetic biology: new engineering rules for an emerging discipline	http://onlinelibrary.wiley.com/doi/10.1038/msb4100073/full									
15	The second wave of synthetic biology: from modules to systems	http://www.nature.com/nrm/journal/v10/n6/full/nrm2698.html									
16	Synthetic biology	http://www.nature.com/nrg/journal/v6/n7/full/nrg1637.html									
17	Synthetic biology: engineering Escherichia coli to see light	http://www.nature.com/nature/journal/v438/n7067/full/nature04405.html									
18	Synthetic biology tools	https://www	Filed Jun 5, 2	Published De	Chunbo LOU	The Regents	WO2012170436A1				
19	Microfluidic platform for synthetic biology applications	https://www	Filed Apr 2, 2	Published Oc	Chieh Chang	The Regents	US20120258487				
20	Synthetic Biology: FAQ	http://syntheticbiology.org/FAQ.html						Feb 28, 2003			
21	Synthetic biology - Wikipedia, the free encyclopedia	https://en.wikipedia.org/wiki/Synthetic_biology						Dec 23, 2015			
22	Regenesis: How Synthetic Biology Will Reinvent Nature and ...	https://books.google.com		2014					George M. Cl	Page 306	No preview
23	Synthetic Biology: Tools and Applications	https://books.google.com		2013						Page 352	Preview
24											
25	Scholar										
26	Title	URL									
27	Synthetic biology: new engineering rules for an emerging discipline	http://onlinelibrary.wiley.com/doi/10.1038/msb4100073/full									
28	The second wave of synthetic biology: from modules to systems	http://www.nature.com/nrm/journal/v10/n6/full/nrm2698.html									
29	Synthetic biology	http://www.nature.com/nrg/journal/v6/n7/full/nrg1637.html									
30	Synthetic biology: engineering Escherichia coli to see light	http://www.nature.com/nature/journal/v438/n7067/full/nature04405.html									
31	Synthetic biology: applications come of age	http://www.nature.com/nrg/journal/v11/n5/full/nrg2775.html									
32	An ER-mitochondria tethering complex revealed by a synthetic biology screen	http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2933203/									
33	Metabolic engineering of microorganisms for biofuels production: from bugs to synthetic	http://download.bioon.com.cn/upload/month_0906/20090602_9baf7ed8bf3db7b1eb2eXd9VaYDZvzAN.attach.pdf									
34	That was the synthetic biology that was	http://thetarrytownmeetings.org/sites/default/files/ThatWasTheSyntheticBiologyThatWas-Campos.pdf									
35	Engineering microbial consortia: a new frontier in synthetic biology	http://macpost.caltech.edu/groups/fha/publications/brenner.pdf									
36	RNA synthetic biology	http://www.nature.com/nbt/journal/v24/n5/full/nbt1208.html									
37											
38	Patents										
39	Title	URL	Application D	Published Da	Inventor	Assignee	Publication Number				
40	Synthetic biology tools	https://www	Filed Jun 5, 2	Published De	Chunbo LOU	The Regents	WO2012170436A1				
41	Microfluidic platform for synthetic biology applications	https://www	Filed Apr 2, 2	Published Oc	Chieh Chang	The Regents	US20120258487				
42	Synthetic biology vectors	https://www	Filed Nov 5, 2	Issued Sep 1(	Peter Woolf	The Regents	US8835160				
43	Synthetic biology tools	https://www.google.com/	Published Jar	Chunbo Lou	The Regents	US20130005590					
44	Novel synthetic biology-based adcc technology	https://www	Filed Jan 8, 2	Published Oc	Chiang J. Li	1 Globe Heal	WO2015106043A3				
45	Modular genomes for synthetic biology and metabolic engineering	https://www	Filed Oct 12,	Published Oc	Robert Holt	Holt Robert A	US20070243617				
46	Modular genomes for synthetic biology and metabolic engineering	https://www	Filed Oct 12,	Published Oc	Robert A Hol	Bc Cancer Ag	WO2007110695A2				
47	Multiscale platform for coordinating cellular activity using ...	https://www	Filed Dec 14,	Published Jul	Arthur PRINC	The Regents	WO2013090818A1				
48	Modular genomes for synthetic biology and metabolic engineering	https://www	Filed Oct 12,	Published Jul	Robert A. Ho	BC Cancer Ag	EP1948803A2				
49	Modular genomes for synthetic biology and metabolic engineering	https://www	Filed Oct 12,	Published Oc	Robert A. Ho	Bc Cancer Ag	CA2625262A1				

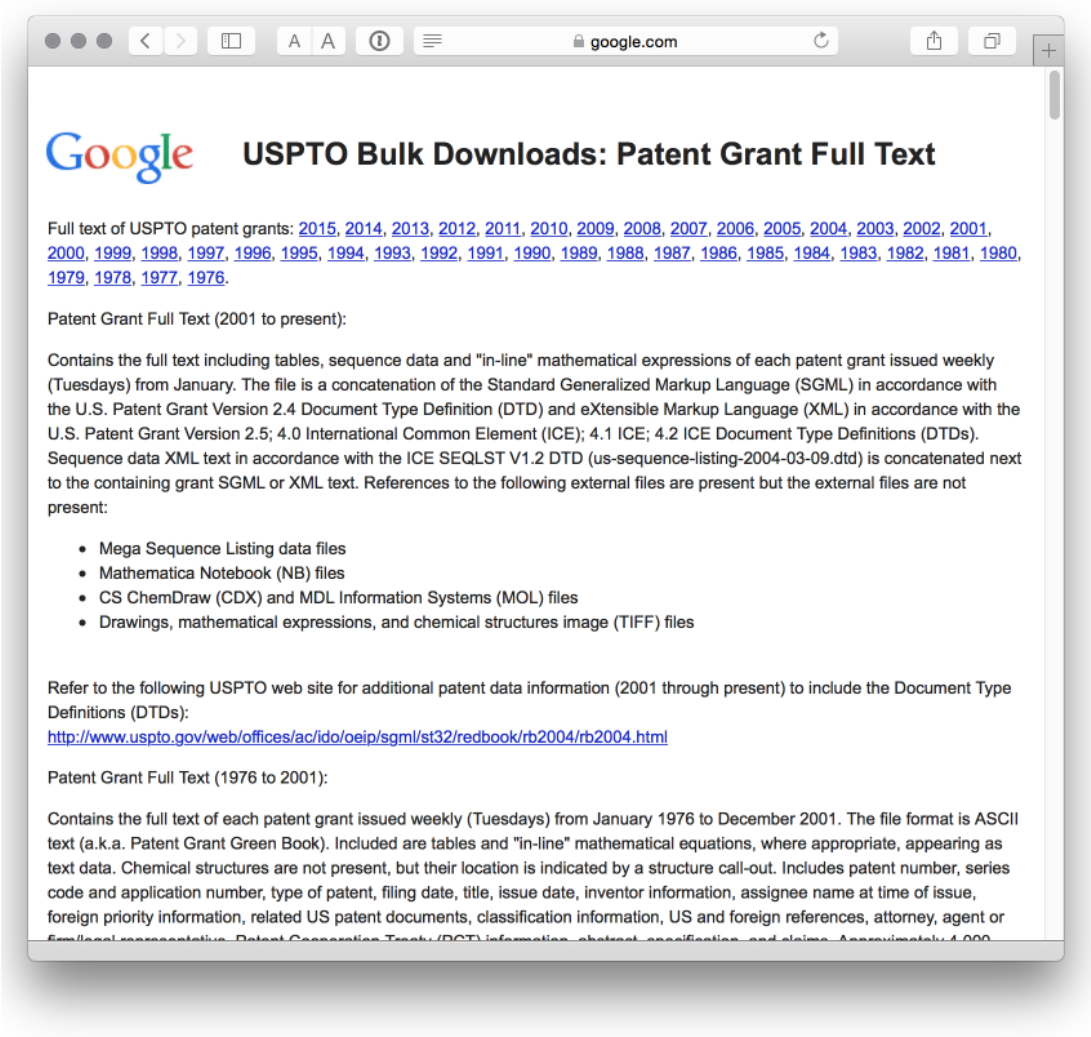
Es posible cargar más resultados para una sección (por ejemplo, ver Más resultados de patentes en la parte inferior de los resultados) y luego exportarlos (por ejemplo, 20 documentos de patentes en lugar de 10). En una prueba logramos exportar 140 resultados de patentes, pero esto podría volverse laborioso rápidamente. Un problema adicional es que los datos necesitarán ser transpuestos. En el momento de redactar este documento, no habíamos identificado una ruta API hacia el Buscador de Arte Anterior.

5.2.9 Descarga masiva de USPTO de Google

Las bases de datos de patentes de la USPTO pueden ser arcaicas, pero puede descargar la colección completa de los EE. UU. Desde el servicio de descarga masiva de la USPTO de Google .



Es un servicio fantástico, y un ejemplo para las oficinas de patentes en todo el mundo en la liberación de datos de patentes. Si tiene una buena conexión de banda ancha y espacio en el disco duro, es bastante divertido tener acceso repentinamente a millones de registros de patentes. Los autores utilizaron el servicio para extraer el texto de la colección de millones de nombres de especies biológicas como se informa aquí .



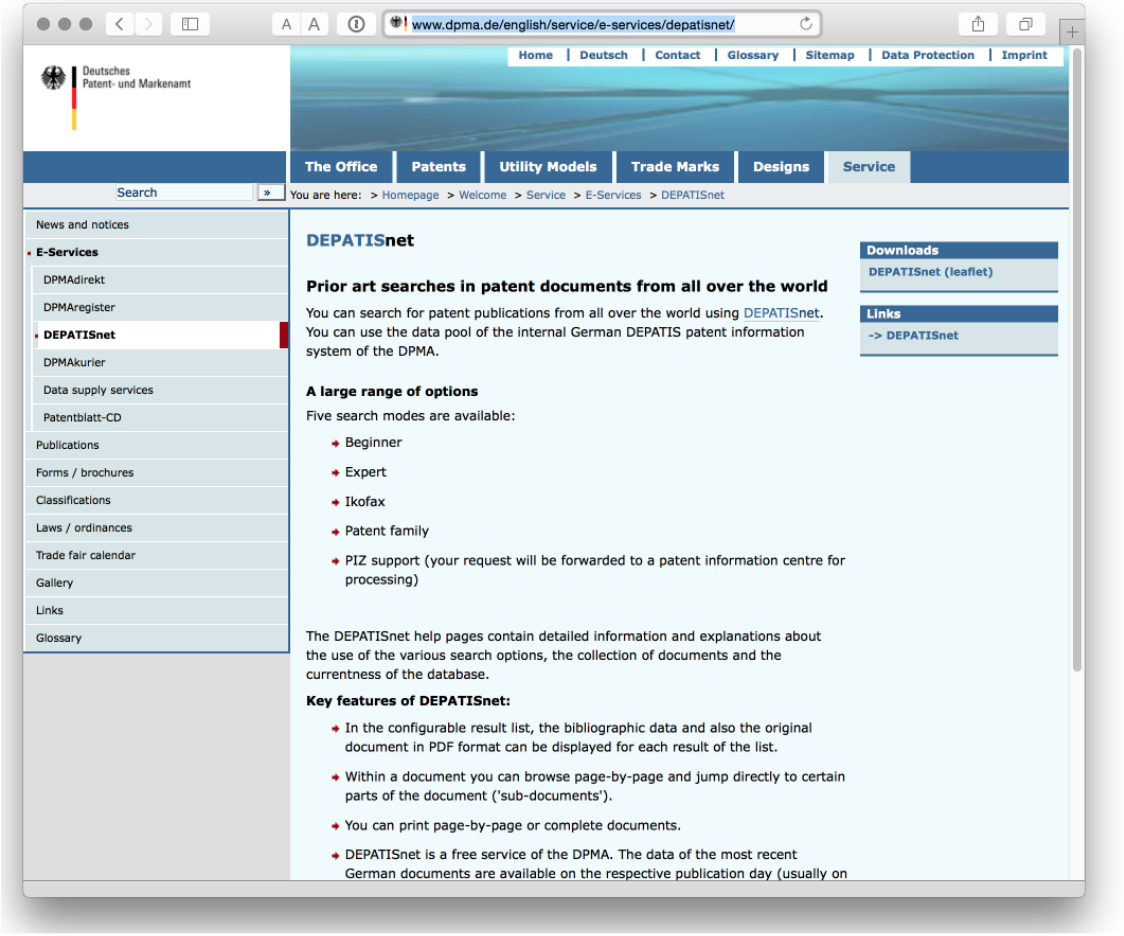
Sin embargo, un aspecto importante a tener en cuenta es que el XML que delimita documentos individuales no siempre está bien demarcado. Esto significa que cualquier código que funcione para un conjunto masivo de archivos puede fallar en otro conjunto. Si bien es posible abordar esto, prepárese para pasar tiempo trabajando en esto y / o busque ayuda de un programador profesional. Para obtener información sobre estos problemas, consulte esta discusión de Stackoverflow sobre el análisis de los datos en R.

5.2.10 Patentes gratis en línea

Regístrese para obtener una cuenta gratuita para un mejor acceso y para guardar y descargar datos. Ha existido desde hace bastante tiempo y aunque las opciones de descarga son limitadas, nos gusta.

5.2.11 DEPATISnet

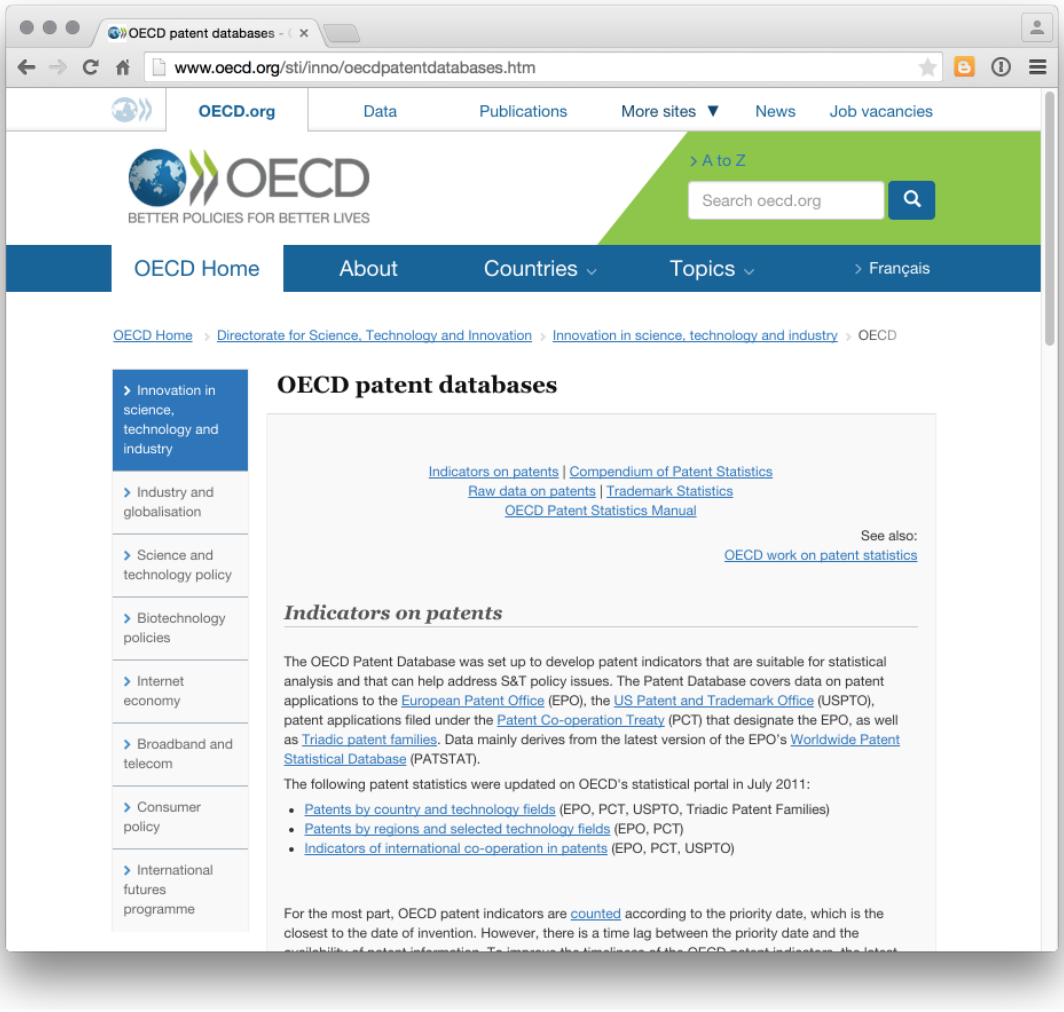
No estamos cubriendo las bases de datos nacionales. Sin embargo, la base de datos de patentes de la Oficina Alemana de Patentes y Marcas nos pareció potencialmente muy útil. Permite búsquedas en inglés y alemán y tiene una amplia cobertura de datos de patentes internacionales, incluidas las colecciones de China, EP, EE. UU. Y PCT. Los detalles de la cobertura están aquí . Vale la pena experimentar con.



5.2.12 Bases de datos de patentes de la OCDE

Una que es más para los estadísticos de patentes. La OCDE ha invertido un gran esfuerzo en el desarrollo de indicadores de patentes y recursos, incluidas citas, la

base de datos HAN de nombres de solicitantes armonizados , la cartografía a través de la base de datos REGPAT, entre otros recursos disponibles de forma gratuita.

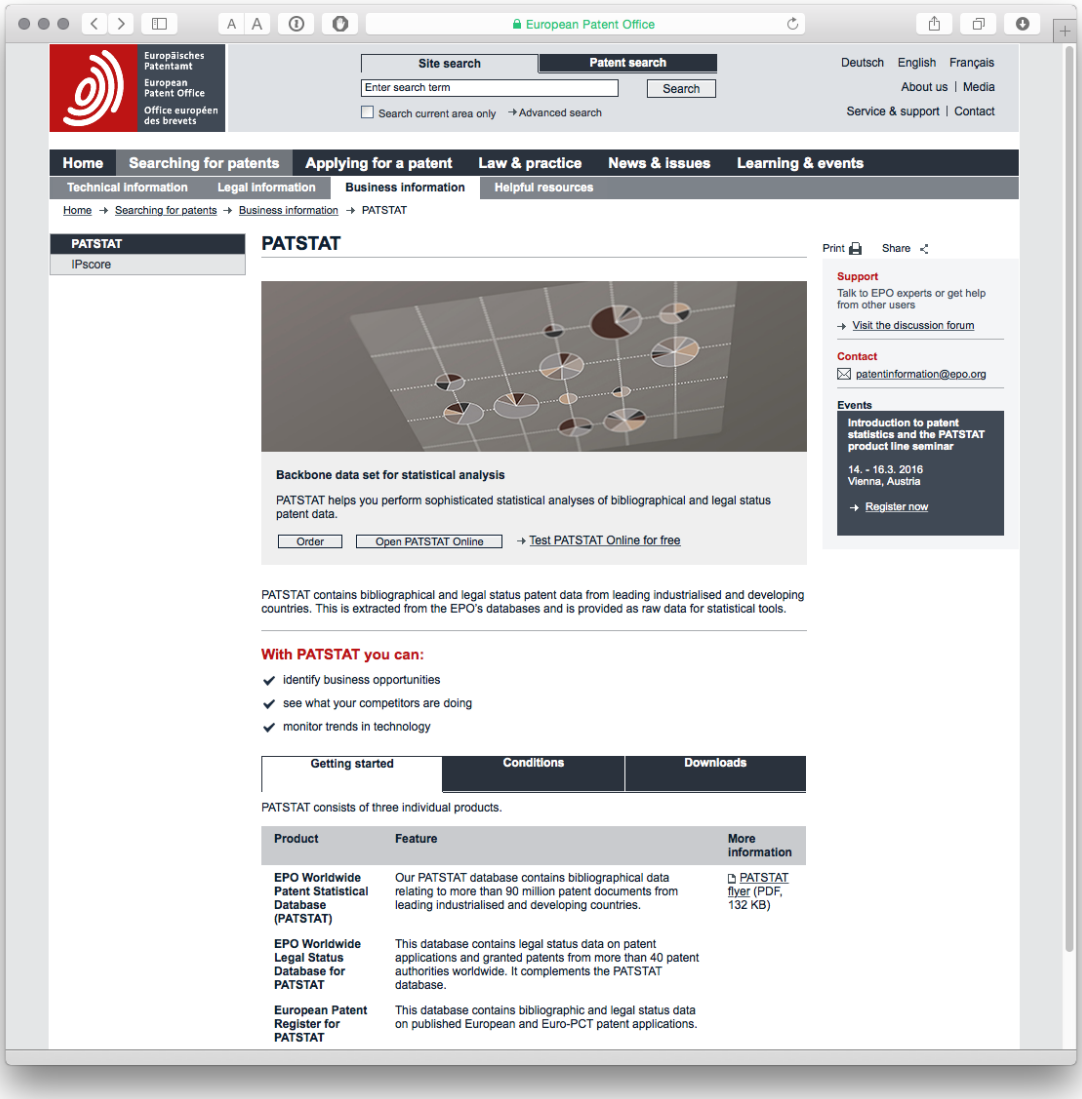


En la misma línea, el archivo de datos de citas de patentes de la Oficina Nacional de Investigación Económica de EE. UU. De EE. UU.Es un recurso importante.

5.2.13 Base de datos estadísticos de patentes mundiales de la OEP

La base de datos más importante para uso estadístico es la Base de Datos Estadísticos de Patentes Mundiales de la OEP (PATSTAT) y contiene alrededor de 90 millones de registros. PATSTAT no es gratis y cuesta 1250 euros por un año (dos ediciones) o 630 euros por una sola edición. La principal barrera para usar PATSTAT es la necesidad de ejecutar y mantener una base de datos de +200 Gigabyte. Sin embargo, también hay una versión en línea de PATSTAT que es

gratuita durante los primeros dos meses si desea probarlo inscribiéndose en la prueba (se requiere conocimiento de SQL).

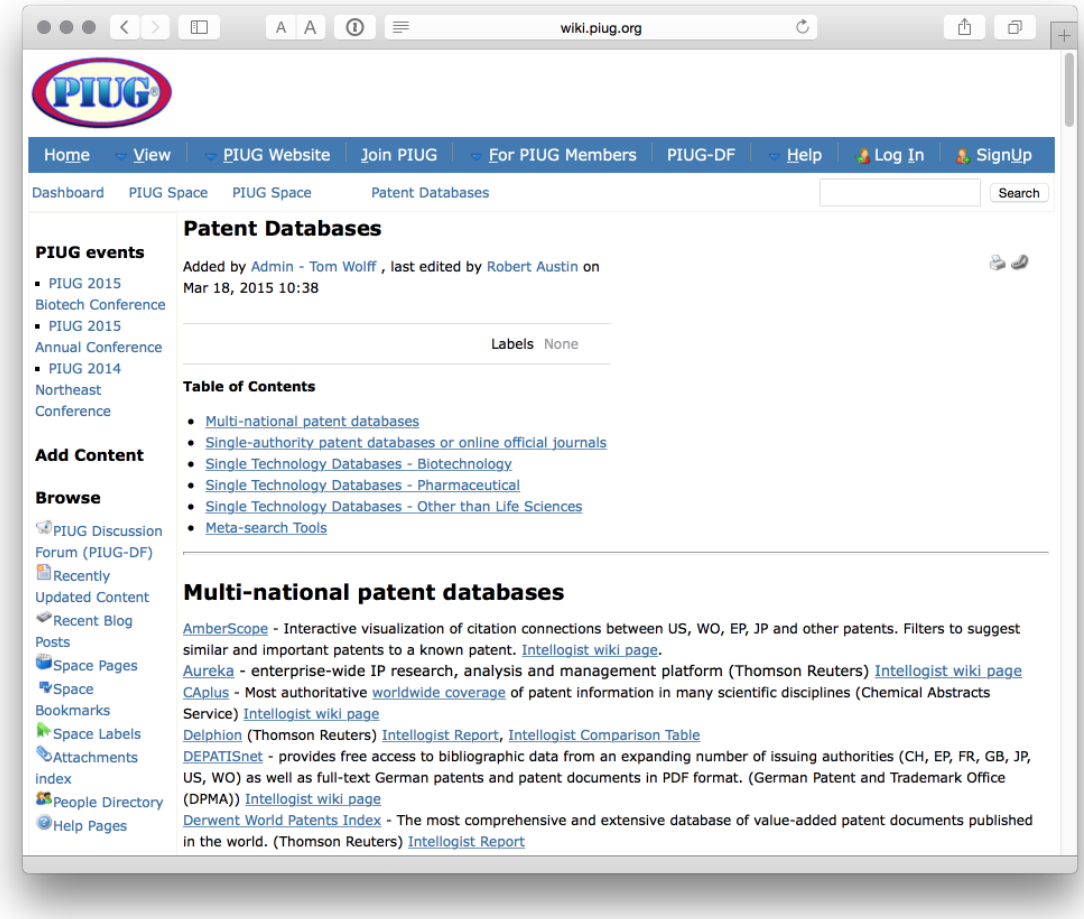


Para los usuarios que buscan cargar PATSTAT en una base de datos MySQL, Simone Mainardi proporciona el siguiente código en Github .

5.2.14 Otras fuentes de datos

Varias compañías brindan acceso a los datos de patentes, por lo general con acceso escalonado según sus necesidades y presupuesto. Los ejemplos incluyen Thomson Innovation , Questel Orbit , STN y PatBase . No nos enfocaremos en estos servicios, pero analizaremos el uso de herramientas de datos para trabajar con datos de servicios como Thomson Innovation.

Para obtener más información sobre proveedores de datos comerciales y gratuitos, pruebe el excelente Grupo de usuarios de información sobre patentes y su lista de bases de datos de patentesde Tom Wolff y Robert Austin.



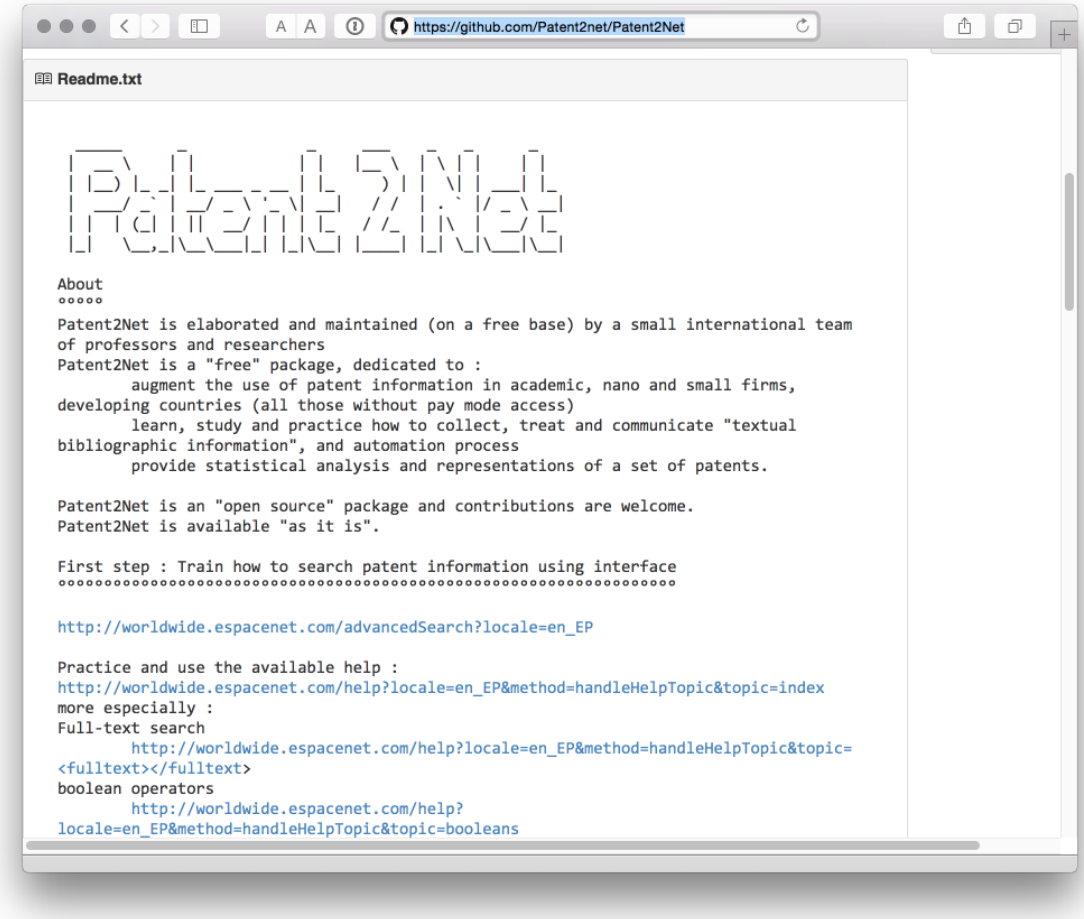
También vale la pena mencionar el blog Landon IP Intellogist que mantiene los informes del sistema de búsqueda.

Herramientas para acceder a datos de patentes

Al cerrar este capítulo, resaltaremos un par de herramientas para acceder a los datos de patentes, generalmente utilizando API y Python. Volveremos sobre esto más adelante y estamos trabajando para probar este enfoque en R.

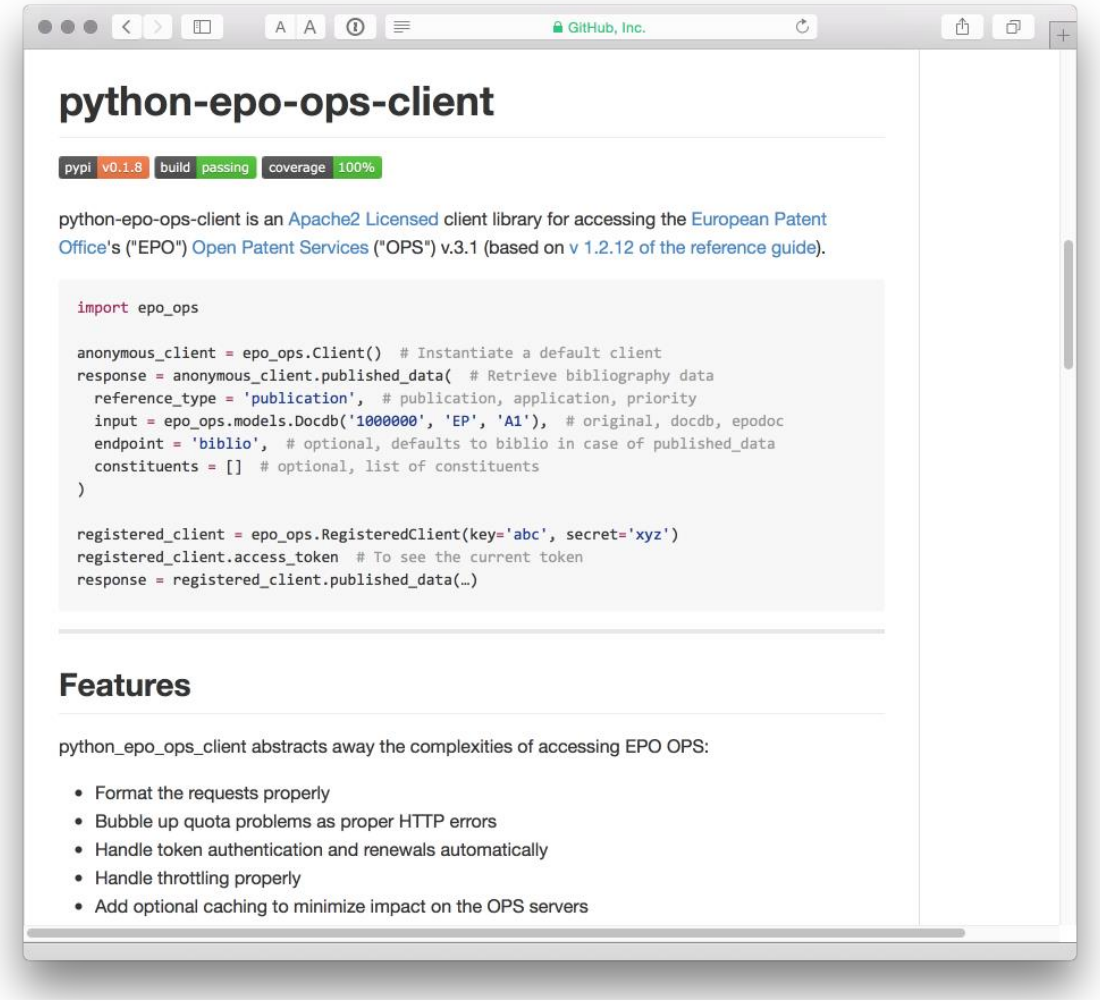
5.2.15 Patent2Net en Python

Una herramienta de Python para acceder y procesar los datos del servicio OPS de la Oficina Europea de Patentes.



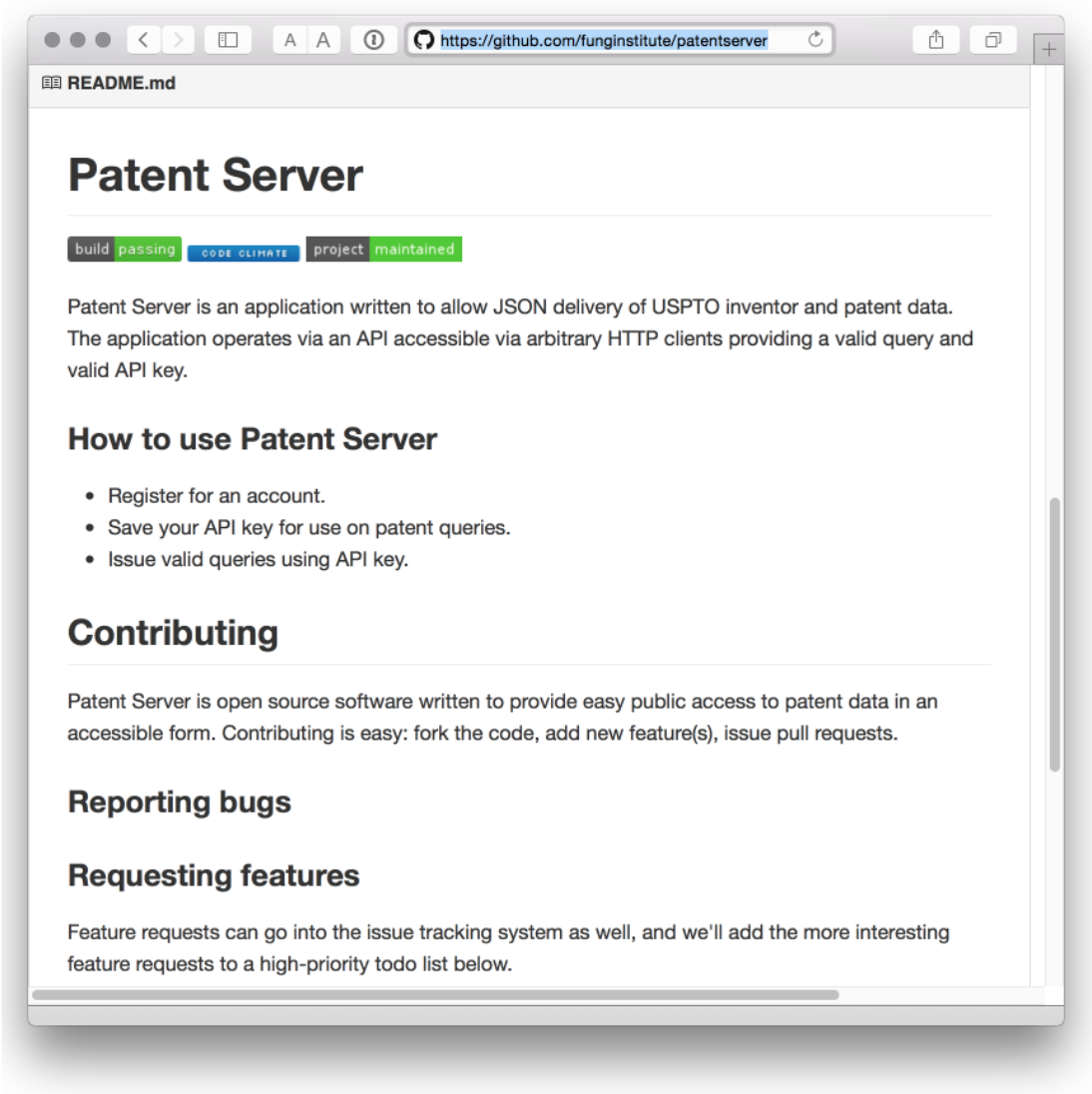
5.2.16 Cliente Python EPO OPS de Gsong

Un cliente de Python para acceso OPS desarrollado por Gsong y disponible gratuitamente en GitHub. Utilizado en Patent2Net arriba.



5.2.17 Fung Institute Patent Server para datos USPTO en JSON

Los investigadores del Instituto Fung también han estado activos en el desarrollo de recursos de código abierto para acceder y trabajar con datos de patentes. Destacamos patentserverpero vale la pena echarle un vistazo a otros recursos en el repositorio como patentprocessor , un conjunto de scripts de Python para el procesamiento de datos de mayor descarga de la USPTO. Tenga en cuenta que el desarrollo de estas herramientas ya no parece estar activo.



##Redondeo

Un problema que enfrentan los analistas de patentes es el acceso a los datos en una forma adecuada para un análisis más detallado. Típicamente esto involucra cientos o muchos miles de registros. Los últimos años han abierto cada vez más los datos de patentes gracias a la posibilidad de descargar 1,000 o 10,000 registros a la vez. Sin embargo, el acceso a descargas de títulos, resúmenes y reclamaciones o descripciones y el texto completo sigue siendo limitado cuando esto es lo que se necesita. Las oficinas de patentes, como la USPTO, han asumido un papel de liderazgo en la disponibilidad de datos de patentes a granel y esto es muy bienvenido para quienes trabajan en el análisis de patentes. Sin embargo, es razonable decir que la situación actual es una de las mejoras en el acceso (a través

de Patentscope, el Lens y el servicio EPO OPS) pero no del todo en las cantidades o con los campos de datos que les gustaría a los analistas de patentes.

Capítulo 6 The Lens

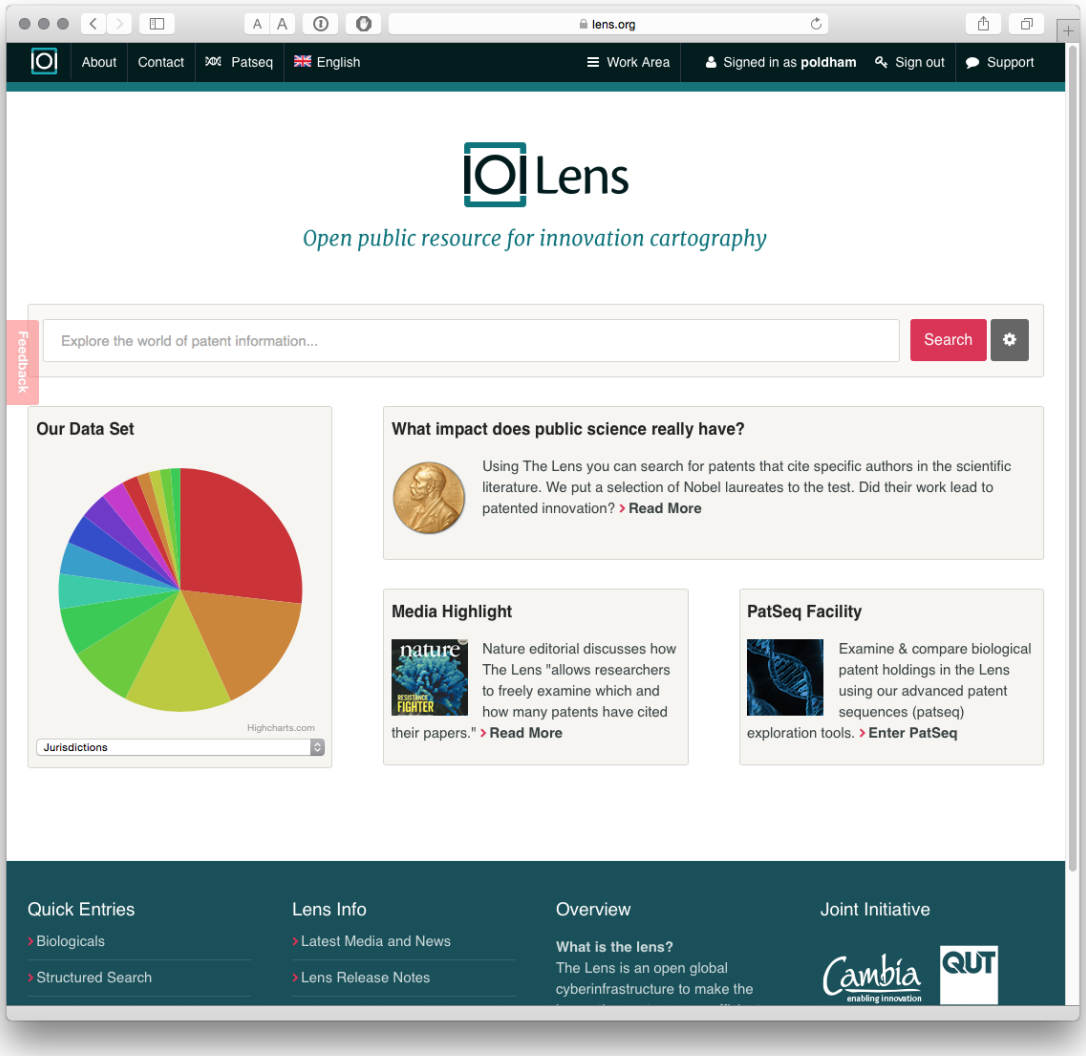
6.1 Introducción

En este capítulo, proporcionamos una breve introducción a la base de datos de patentes de The Lens como fuente gratuita de datos para el análisis de patentes.

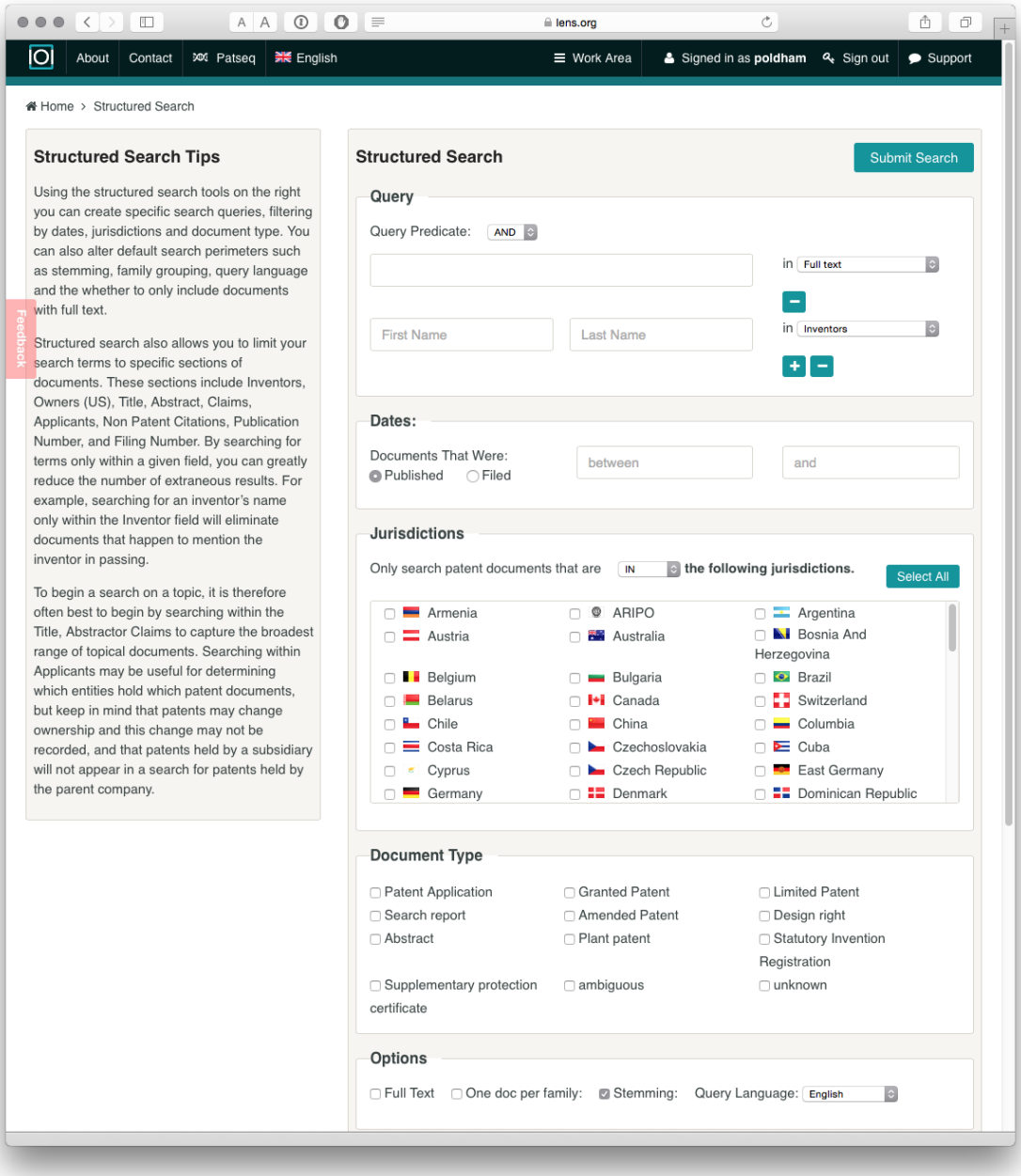
The Lens es una base de datos de patentes con sede en Australia que se describe a sí misma como "una ciberinfraestructura global abierta para hacer que el sistema de innovación sea más eficiente y justo, más transparente e inclusivo". La principal forma de hacerlo es proporcionar acceso a la información de patentes con un enfoque particular en la información de secuencia, así como el análisis de temas como la actividad de patentes relacionada con el ADN. Una característica importante de The Lens para quienes trabajan en temas relacionados con la biotecnología es PatSeq .

6.2 Primeros pasos

Para aprovechar al máximo el objetivo, el primer paso es registrarse para obtener una cuenta desde la página principal.



Es posible comenzar a buscar directamente desde la página principal. Sin embargo, al seleccionar el botón pequeño junto al cuadro de búsqueda, se accede a los controles de búsqueda.



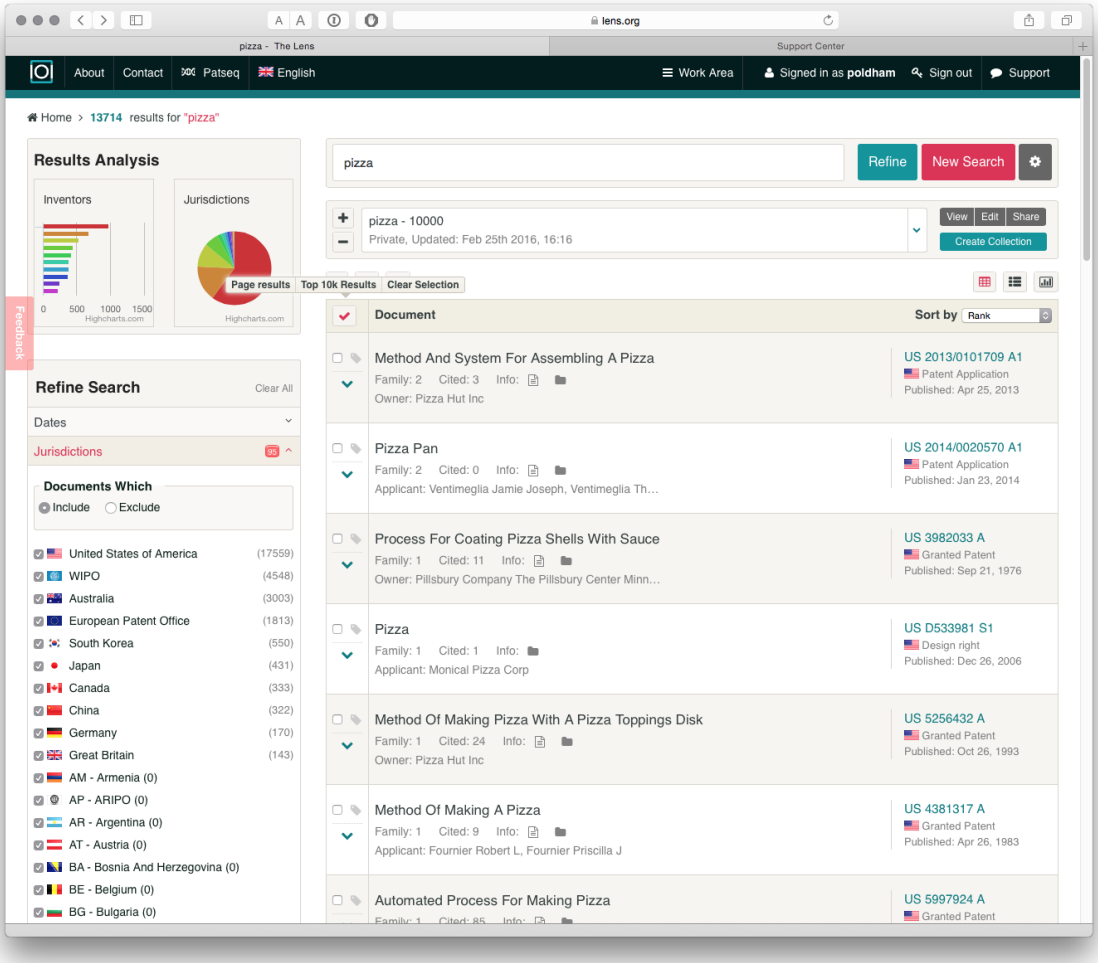
Como podemos ver, podemos usar consultas booleanas para buscar en una variedad de campos que incluyen el texto completo, el título, el resumen o las reclamaciones (una ventaja importante). También podemos seleccionar una o varias jurisdicciones. Además, los resultados se pueden refinar a solicitudes de patentes o subvenciones, y hay opciones para el texto completo o un documento por familia (lo que reduce considerablemente el número de resultados).

Utilizamos nuestra consulta estándar "pizza", todas las jurisdicciones y un documento por familia. Nos dimos vuelta parando.

Nuestra búsqueda de pizza arrojó 13,714 familias de un total de 29,617 publicaciones que contienen el término en el texto completo. Este enfoque ayuda a refinar las búsquedas al reducir la duplicación.

Lens permite a los usuarios crear colecciones de hasta 10,000 resultados de una búsqueda. Para crear una colección use el Create Collectionbotón y nombre la colección. La forma en que agrega registros a una colección no es obvia e implica 2 pasos.

1. Verifique la flecha al lado de Documento como en la imagen de abajo. Cuando el mouse se desplace sobre la flecha, aparecerá un menú emergente. Elija Top 10k Results.
2. En el cuadro que muestra el nombre de la colección sobre los resultados, presione la flecha + para agregar los 10,000 documentos a la Colección.



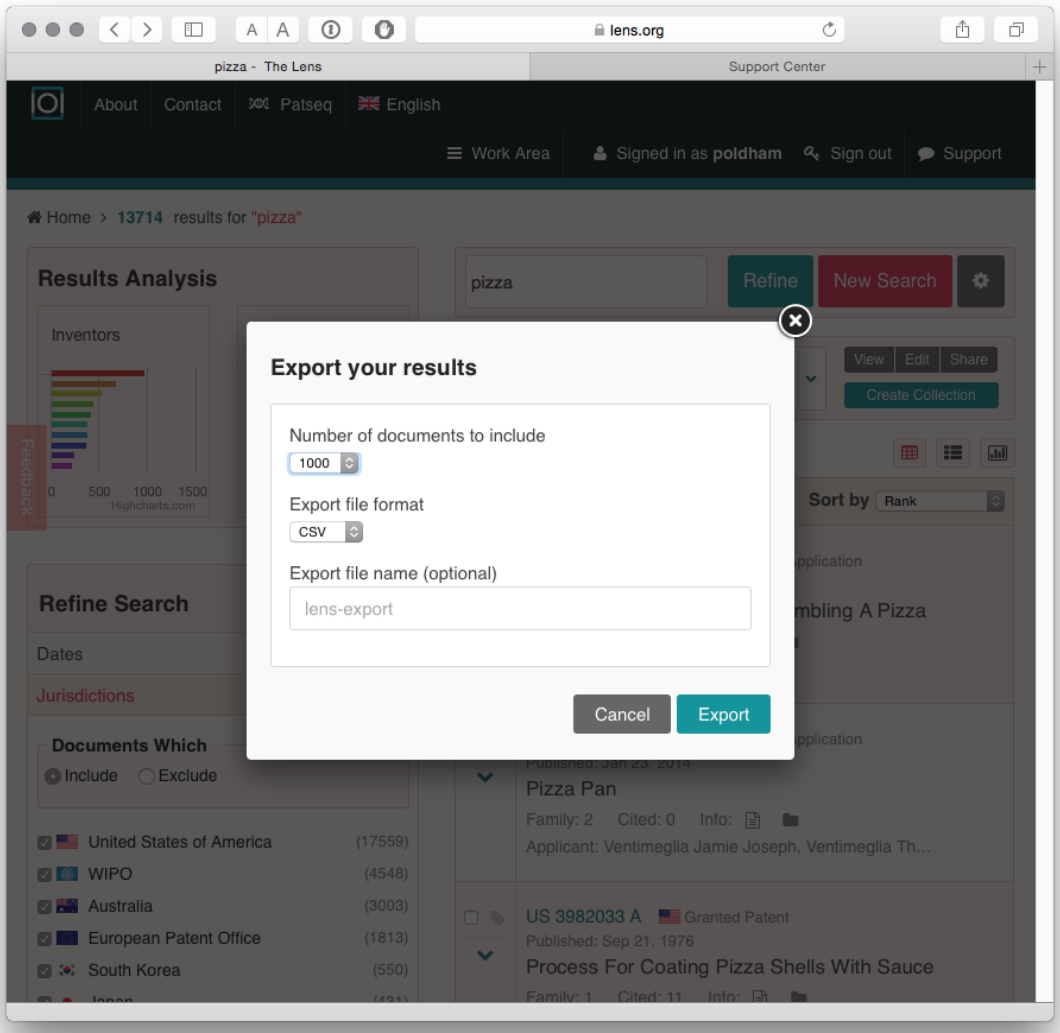
Una vez que entienda este proceso, es fácil agregar documentos a las colecciones. Una muy buena característica de la lens es que cuando se crea una colección, podemos compartirla con otras personas mediante el Sharebotón. Los usuarios tienen la opción de mantener una colección privada o compartir públicamente. La URL de la colección que acabamos de generar es https://www.lens.org/lens/collection/9606 .

Podríamos imaginar que, para búsquedas más restringidas, y teniendo en cuenta las cuestiones de confidencialidad, esta podría ser una forma útil de compartir datos de patentes con colegas. Una adición útil sería la posibilidad de compartir con grupos basados en direcciones de correo electrónico o algo similar (aunque eso puede ser posible al elegir un enlace privado y compartirlo).

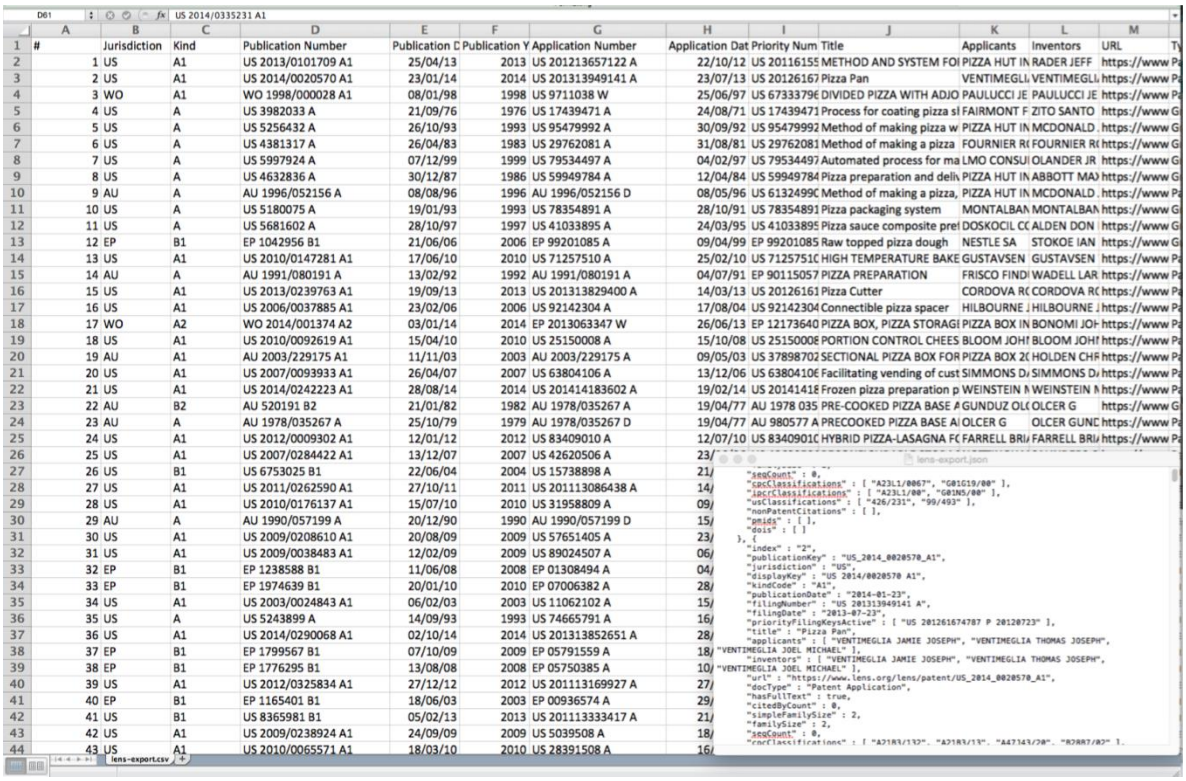
Al usar los pequeños íconos de arriba Documenta la izquierda, podemos guardar nuestra consulta para su uso posterior, limitar los datos a familias simples o expandir a publicaciones, y descargar los datos.

Hay dos opciones principales para descargar datos. El primero es descargar 1000 registros seleccionando el botón de exportación arriba Document.

Cuando seleccionamos el botón de exportación, se nos presentará una opción sobre el número de registros a exportar y si exportamos en JSON (para uso programático), RIS para software bibliográfico o .csv para usar en herramientas como Excel u otros programas.



Los resultados de la exportación son claros y claros sobre lo que representan en comparación con algunas bases de datos de patentes. url También se proporciona un enlace al archivo relevante en la Lens que puede ayudar en la revisión de documentos.

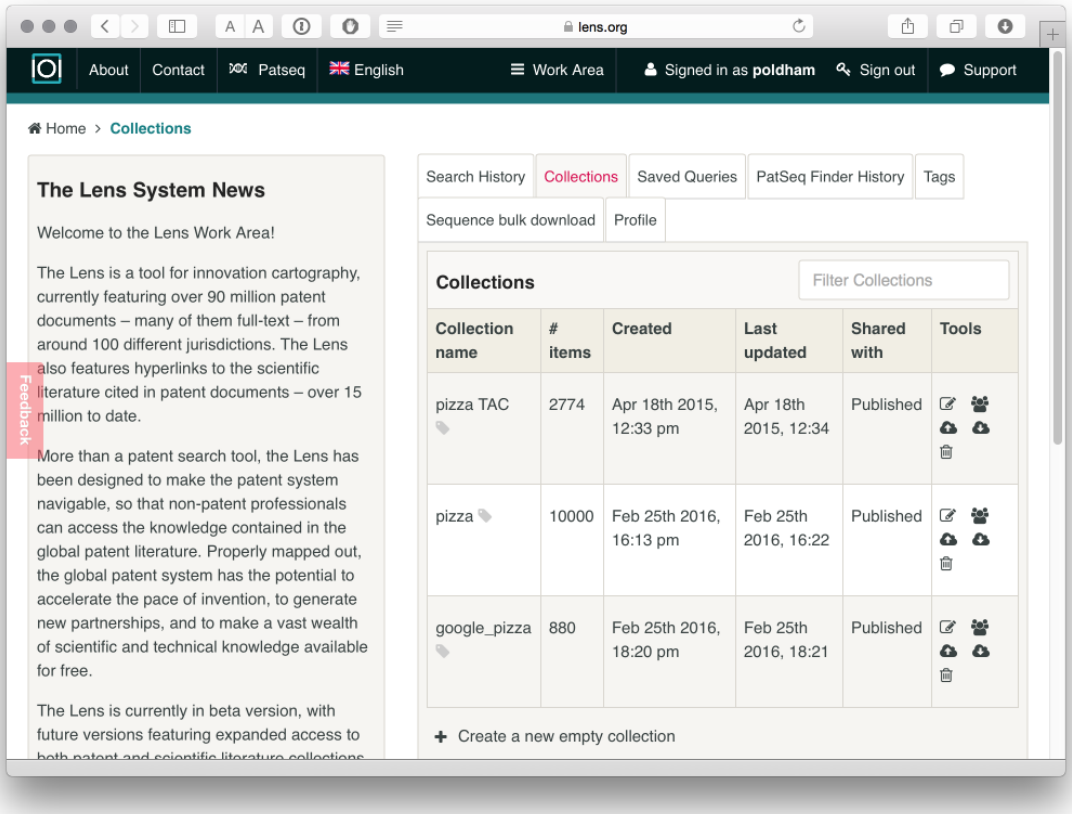


#	Jurisdiction	Kind	Publication Number	Publication D	Publication Y	Application Number	Application Dat	Priority Num	Title	Applicants	Inventors	URL
1	US	A1	US 2013/0101709 A1	25/04/13	2013	US 201213657122 A	22/10/12	US 20116155	METHOD AND SYSTEM FOI	PIZZA HUT IN	RADER JEFF	https://www
2	US	A1	US 2014/0020570 A1	23/01/14	2014	US 201313949141 A	23/07/13	US 20126167	Pizza Pan	VENTIMEGLI	VENTIMEGLI	https://www
3	WO	A1	WO 1998/000028 A1	08/01/98	1998	US 9711038 W	25/06/97	US 6733379	DIVIDED PIZZA WITH ADJO	PAULUCCI JE	PAULUCCI JE	https://www
4	US	A	US 3982033 A	21/09/76	1976	US 17439471 A	24/08/71	US 17439471	Process for coating pizza sl	FAIRMONT F	ZITO SANTO	https://www
5	US	A	US 5256432 A	26/10/93	1993	US 95479992 A	30/09/92	US 95479992	Method of making pizza w	PIZZA HUT IN	MCDONALD	https://www
6	US	A	US 4381317 A	26/04/83	1983	US 29762081 A	31/08/81	US 29762081	Method of making a pizza	FOURNIER R(	FOURNIER R(	https://www
7	US	A	US 5997924 A	07/12/99	1999	US 79534497 A	04/02/97	US 79534497	Automated process for ma	LMO CONSUI	OLANDER JR	https://www
8	US	A	US 4632836 A	30/12/87	1986	US 59949784 A	12/04/84	US 59949784	Pizza preparation and deli	PIZZA HUT IN	ABBOTT MA	https://www
9	AU	A	AU 1996/052156 A	08/08/96	1996	AU 1996/052156 D	08/05/96	US 6132499	Method of making a pizza,	PIZZA HUT IN	MCDONALD	https://www
10	US	A	US 5180075 A	19/01/93	1993	US 78354891 A	28/10/91	US 78354891	Pizza packaging system	MONTALBAN	MONTALBAN	https://www
11	US	A	US 5681602 A	28/10/97	1997	US 41033895 A	24/03/95	US 41033895	Pizza sauce composite pre	DOSKOCIL CC	ALDEN DON	https://www
12	EP	B1	EP 1042956 B1	21/06/06	2006	EP 99201085 A	09/04/99	EP 99201085	Raw topped pizza dough	NESTLE SA	STOKOE IAN	https://www
13	US	A1	US 2010/0147281 A1	17/06/10	2010	US 71257510 A	25/02/10	US 7125751(	HIGH TEMPERATURE BAKE	GUSTAVSEN	GUSTAVSEN	https://www
14	AU	A	AU 1991/080191 A	13/02/92	1992	AU 1991/080191 A	04/07/91	EP 90115057	PIZZA PREPARATION	FRISCO FIND	WADELL LAR	https://www
15	US	A1	US 2013/0239763 A1	19/09/13	2013	US 201313829400 A	14/03/13	US 20126161	Pizza Cutter	CORDOVA R(	CORDOVA R(	https://www
16	US	A1	US 2006/0037885 A1	23/02/06	2006	US 92142304 A	17/08/04	US 9214230	Connectible pizza spacer	HILBOURNE J	HILBOURNE J	https://www
17	WO	A2	WO 2014/001374 A2	03/01/14	2014	EP 12173640 W	26/06/13	EP 12173640	PIZZA BOX, PIZZA STORAGI	PIZZA BOX IN	BONOMI JOH	https://www
18	US	A1	US 2010/0092619 A1	15/04/10	2010	US 25150008 A	15/10/08	US 2515000	PORTION CONTROL CHEES	BLOOM JOH	BLOOM JOH	https://www
19	AU	A1	AU 2003/229175 A1	11/11/03	2003	AU 2003/229175 A	09/05/03	US 37898702	SECTIONAL PIZZA BOX FOR	PIZZA BOX 2(	HOLDEN CHF	https://www
20	US	A1	US 2007/0093933 A1	26/04/07	2007	US 63804106 A	13/12/06	US 6380410(	Facilitating vending of cust	SIMMONS D/	SIMMONS D/	https://www
21	US	A1	US 2014/0242223 A1	28/08/14	2014	US 201414183602 A	19/02/14	US 2014141	Frozen pizza preparation p	WEINSTEIN M	WEINSTEIN M	https://www
22	AU	B2	AU 520191 B2	21/01/82	1982	AU 1978/035267 A	19/04/77	AU 1978 035	PRE-COOKED PIZZA BASE A	GUNDUZ OL(	OLCER G	https://www
23	AU	A	AU 1978/035267 A	25/10/79	1979	AU 1978/035267 D	19/04/77	AU 980577 A	PRECOOKED PIZZA BASE A	OLCER G	OLCER GUNC	https://www
24	US	A1	US 2012/0009302 A1	12/01/12	2012	US 83409010 A	12/07/10	US 8340901(	HYBRID PIZZA-LASAGNA FC	FARRELL BRI	FARRELL BRI	https://www
25	US	A1	US 2007/0284422 A1	13/12/07	2007	US 42620506 A	23/					
26	US	B1	US 6753025 B1	22/06/04	2004	US 15738898 A	21/					
27	US	A1	US 2011/0262590 A1	27/10/11	2011	US 201113086438 A	14/					
28	US	A1	US 2010/0176137 A1	15/07/10	2010	US 31958809 A	09/					
29	AU	A	AU 1990/057199 A	20/12/90	1990	AU 1990/057199 D	15/					
30	US	A1	US 2009/0208610 A1	20/08/09	2009	US 57651405 A	23/					
31	US	A1	US 2009/0038483 A1	12/02/09	2009	US 89024507 A	06/					
32	EP	B1	EP 1238588 B1	11/06/08	2008	EP 01308494 A	04/					
33	EP	B1	EP 1974639 B1	20/01/10	2010	EP 07006382 A	28/					
34	US	A1	US 2003/0024843 A1	06/02/03	2003	US 11062102 A	15/					
35	US	A	US 5243899 A	14/09/93	1993	US 74665791 A	16/					
36	US	A1	US 2014/0290068 A1	02/10/14	2014	US 201313852651 A	28/					
37	EP	B1	EP 1799567 B1	07/10/09	2009	EP 05791559 A	18/					
38	EP	B1	EP 1776295 B1	13/08/08	2008	EP 05750385 A	10/					
39	US	A1	US 2012/0325834 A1	27/12/12	2012	US 201113169927 A	27/					
40	EP	B1	EP 1165401 B1	18/06/03	2003	EP 00936574 A	29/					
41	US	B1	US 8365981 B1	05/02/13	2013	US 201113333417 A	21/					
42	US	A1	US 2009/0238924 A1	24/09/09	2009	US 5039508 A	18/					
43	US	A1	US 2010/0065571 A1	18/03/10	2010	US 28391508 A	16/					


```
"seqCount" : 0,
"cpcClassifications" : [ "A23L1/0067", "G01G19/00" ],
"ipcrClassifications" : [ "A23L1/00", "G01N5/00" ],
"usClassifications" : [ "426/231", "99/493" ],
"nonPatentCitations" : [ ],
"pmids" : [ ],
"dois" : [ ]
}, {
"index" : "2",
"publicationKey" : "US_2014_0020570_A1",
"jurisdiction" : "US",
"displayKey" : "US 2014/0020570 A1",
"kindCode" : "A1",
"publicationDate" : "2014-01-23",
"filingNumber" : "US 201313949141 A",
"filingDate" : "2013-07-23",
"priorityFilingKeysActive" : [ "US 201261674787 P 20120723" ],
"title" : "Pizza Pan",
"applicants" : [ "VENTIMEGLIA JAMIE JOSEPH", "VENTIMEGLIA THOMAS JOSEPH",
"VENTIMEGLIA JOEL MICHAEL" ],
"inventors" : [ "VENTIMEGLIA JAMIE JOSEPH", "VENTIMEGLIA THOMAS JOSEPH",
"VENTIMEGLIA JOEL MICHAEL" ],
"url" : "https://www.lens.org/lens/patent/US_2014_0020570_A1",
"docType" : "Patent Application",
"hasFullText" : true,
"citedByCount" : 0,
"simpleFamilySize" : 2,
"familySize" : 2,
"seqCount" : 0,
```


La salida JSON (en la parte inferior derecha de la imagen de arriba) también es agradable y limpia.

La segunda ruta para exportar datos es descargar hasta 10,000 resultados usando las colecciones. Cuando seleccionamos el Work Areaícono en la parte superior de la pantalla y seleccionamos Collections, veremos una nueva pantalla con un rango de íconos junto a una colección individual.



Cuando seleccionamos el ícono de descarga, ahora podemos descargar los 10,000 registros de la colección en formatos .csv, ris o JSON. Esto es muy fácil de usar una vez que entienda cómo navegar por la interfaz.

También tenemos la opción de cargar documentos en una colección usando el botón de carga y luego ingresar identificadores separados por comas. Sin embargo, en el momento de escribir esto no pudimos hacer que esta función tan útil funcionara.

6.3 características adicionales

Además de estas características, también es importante tener en cuenta que las exportaciones de datos incluyen un recuento citado que cuenta el número de registros de patentes / no patentes cited del solicitante.

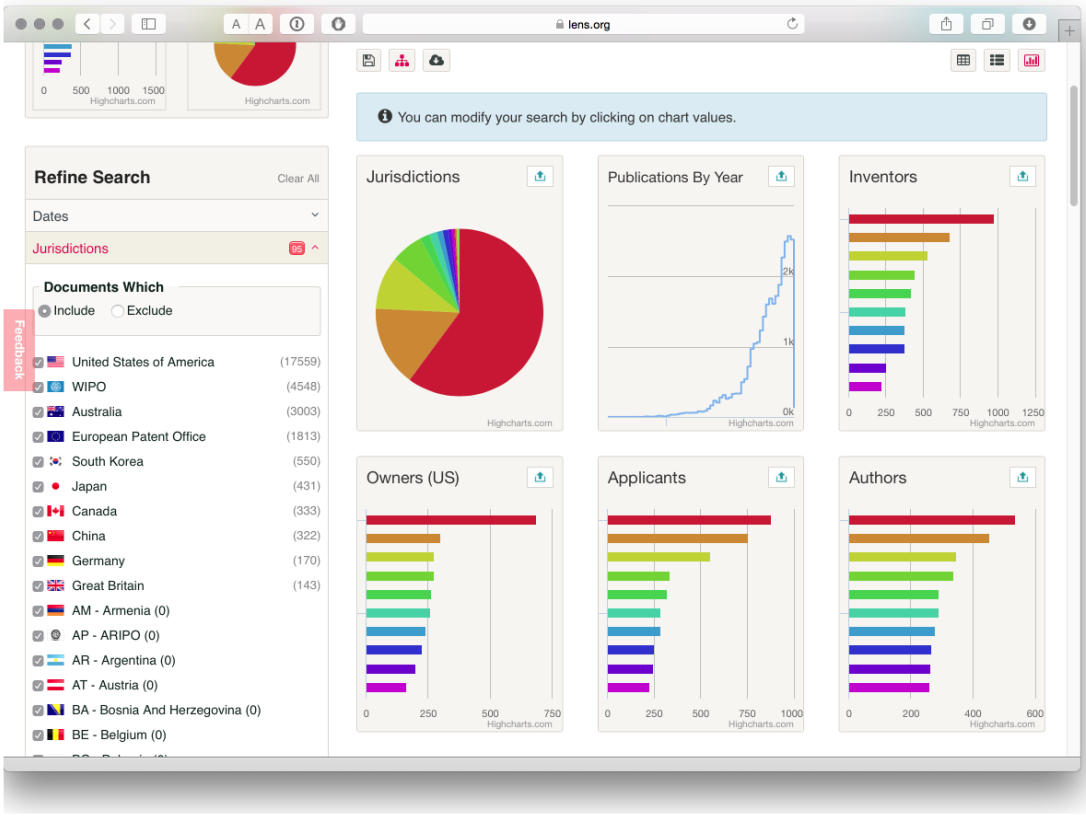
Los datos en línea también muestran los documentos citando. Por ejemplo, el documento US 3982033 A Process for Coating Pizza Shells With Sauce cita tres documentos de patente, pero tiene 11 referencias de futuros postulantes .

Si bien los documentos citados no se incluyen con los datos descargados, es posible visitar un registro de interés en línea y luego crear un nuevo conjunto con los documentos citados. Cuando se han identificado varios documentos de interés, esta podría ser la base para crear una nueva colección de literatura citada o de citas sobre un tema de interés vinculado a una consulta central.

Como tal, un posible flujo de trabajo que utilice The Lens implicaría consultas exploratorias iniciales y refinamiento, descargando los resultados de una consulta refinada para una inspección más detallada y luego seleccionando documentos de interés para explorar las citas atrasadas (citadas) y hacia adelante (citando) y generar una nueva conjunto de datos

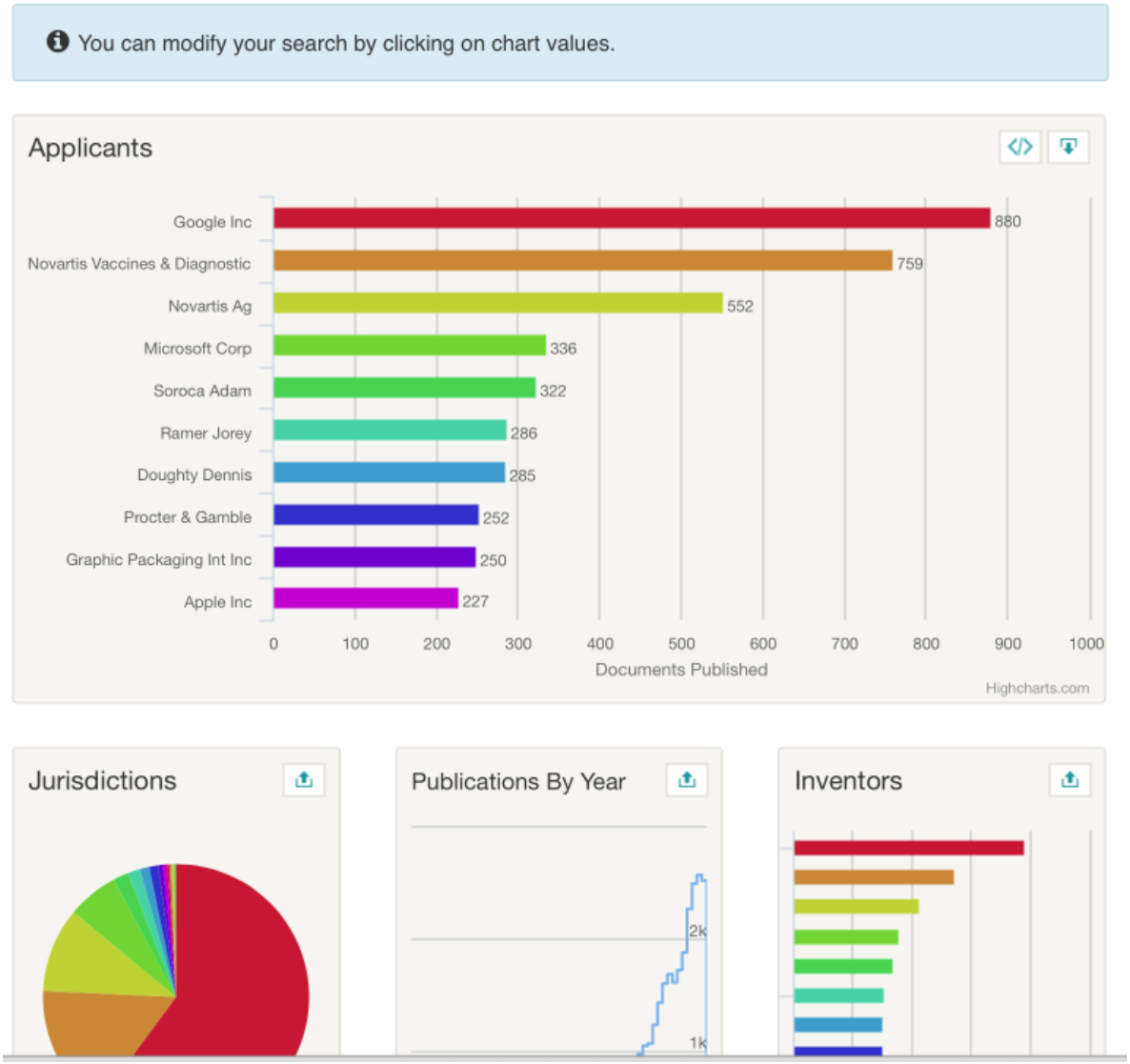
6.4 Visualización

The Lens hace un buen uso de las opciones de visualización en línea usando Highcharts y HTML5. Para acceder a las visualizaciones, elija el pequeño icono a la derecha sobre el Sort bymenú desplegable.

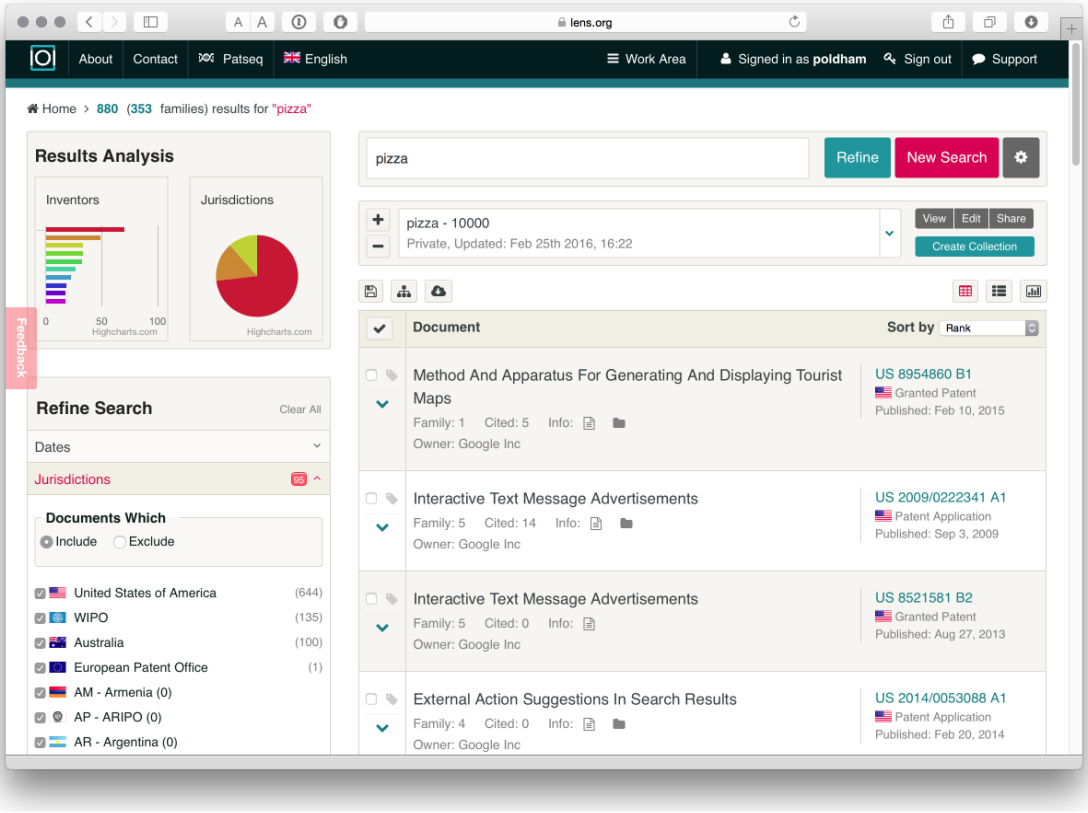


Ahora vemos un conjunto de gráficos para nuestros resultados. Usando el ícono hacia arriba en la parte superior derecha de cada imagen, podemos obtener una

vista ampliada y trabajar con los gráficos. The Lens utiliza la biblioteca de Javascript de Highcharts y una característica muy interesante de este enfoque es que los elementos visuales son interactivos y se pueden usar para refinar los resultados de búsqueda. En la imagen de abajo hemos abierto la imagen de los solicitantes. Como nota aparte, tenga en cuenta que cada imagen se puede copiar como un iframe para incrustar en su propia página web.

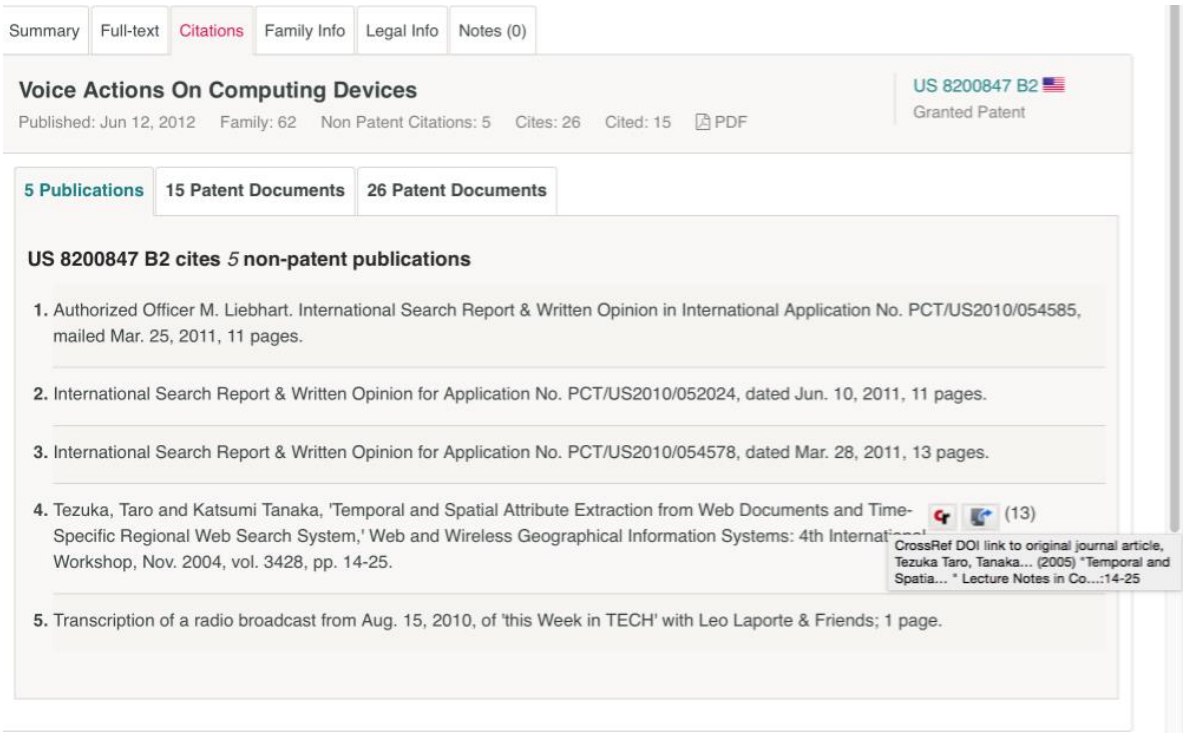


Esto sugiere que Google es el principal usuario de la palabra pizza en el sistema de patentes con 880 documentos en 353 familias. Luego podemos seleccionar el resultado superior y los gráficos se regenerarán enfocándose en nuestra selección (en este caso, Google). Para ver los resultados, necesitamos seleccionar el botón de resultados (el primero a la derecha sobre los cuadros) para ver lo siguiente.



Lo que es muy útil es que es fácil crear una nueva colección para un solicitante de interés, para descargar los resultados o seleccionar áreas de una cartera en función de una jurisdicción o área de tecnología o para explorar patentes altamente citadas. En resumen, podemos profundizar fácilmente en los datos.

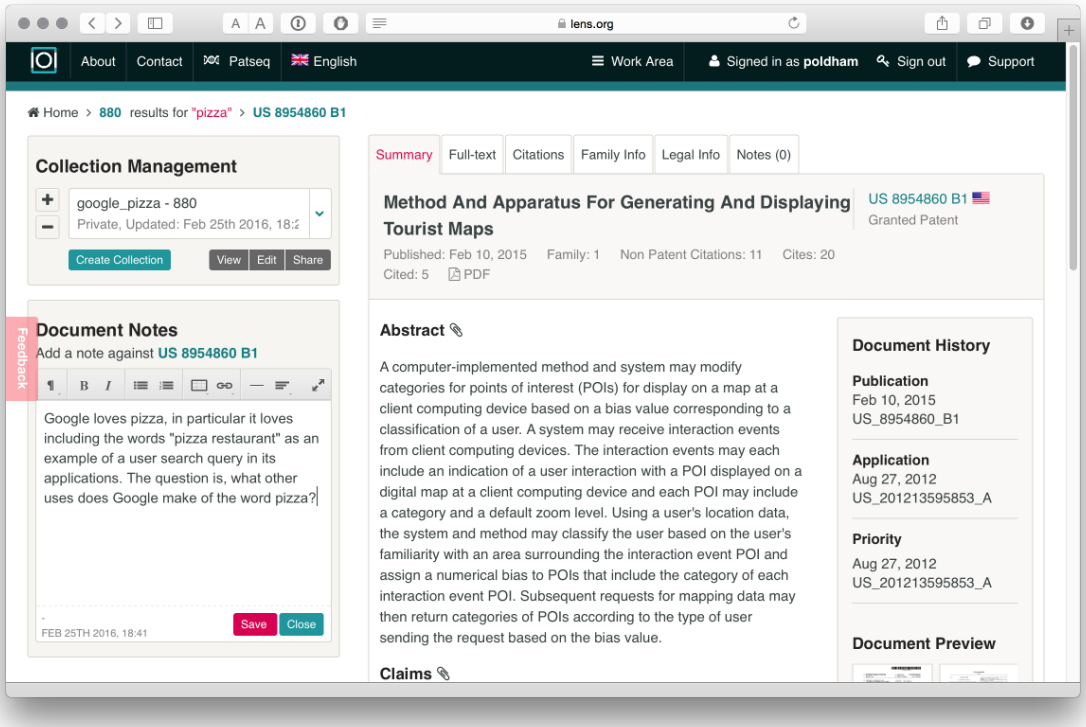
Otras características interesantes del área del gráfico son las referencias a autores, DOI y Id. De PubMed para la exploración de datos extraídos de los documentos. Esto refleja el interés de Lens por investigar la relación entre la investigación científica básica y la innovación. Para acceder a la información relacionada con la literatura, es necesario abrir un gráfico (por ejemplo, autores) y seleccionar el resultado superior y pasar a la vista de resultados. Luego seleccionamos uno de los resultados, como las acciones de voz en los dispositivos informáticos y la pestaña Citas. Esto revela una publicación de un taller sobre sistemas de información geográfica inalámbrica de 2003, como podemos ver a continuación.



Una característica impresionante de este enfoque es el esfuerzo realizado para vincular los datos de las citas con la publicación mediante el uso de referencia cruzada . Según la documentación, alrededor de 15 millones de citas bibliográficas no relacionadas con patentes se han vinculado hasta el momento. Tenga en cuenta que una característica adicional de los datos de descarga de Lens es que incluye un campo de citas bibliográficas que no son patentes. Por ejemplo, la descarga de la cartera de Google Pizza y la búsqueda de la cita anterior revelarán la cita pero sin el valor agregado del DOI. Como tal, la descarga proporcionó los datos NPL en bruto.

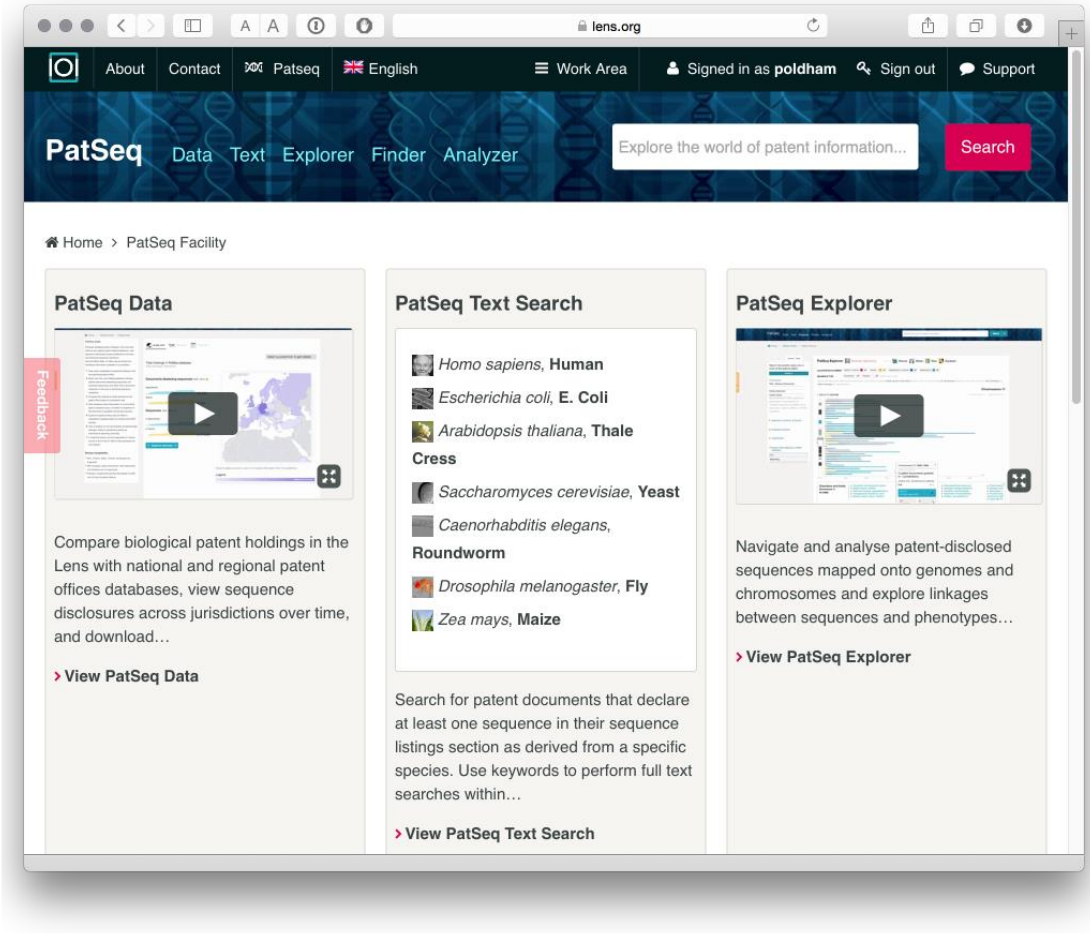
6.5 Trabajando con textos

Al igual que otras bases de datos gratuitas, el objetivo no está diseñado para permitir descargas de múltiples textos completos. Sin embargo, puede acceder al texto completo de los documentos, incluidos los archivos .pdf, y puede tomar notas que se almacenarán con una colección en su cuenta. La imagen a continuación proporciona un ejemplo de nuestros esfuerzos continuos para comprender por qué Google es tan dominante en los resultados de las búsquedas de pizza en documentos de patente.



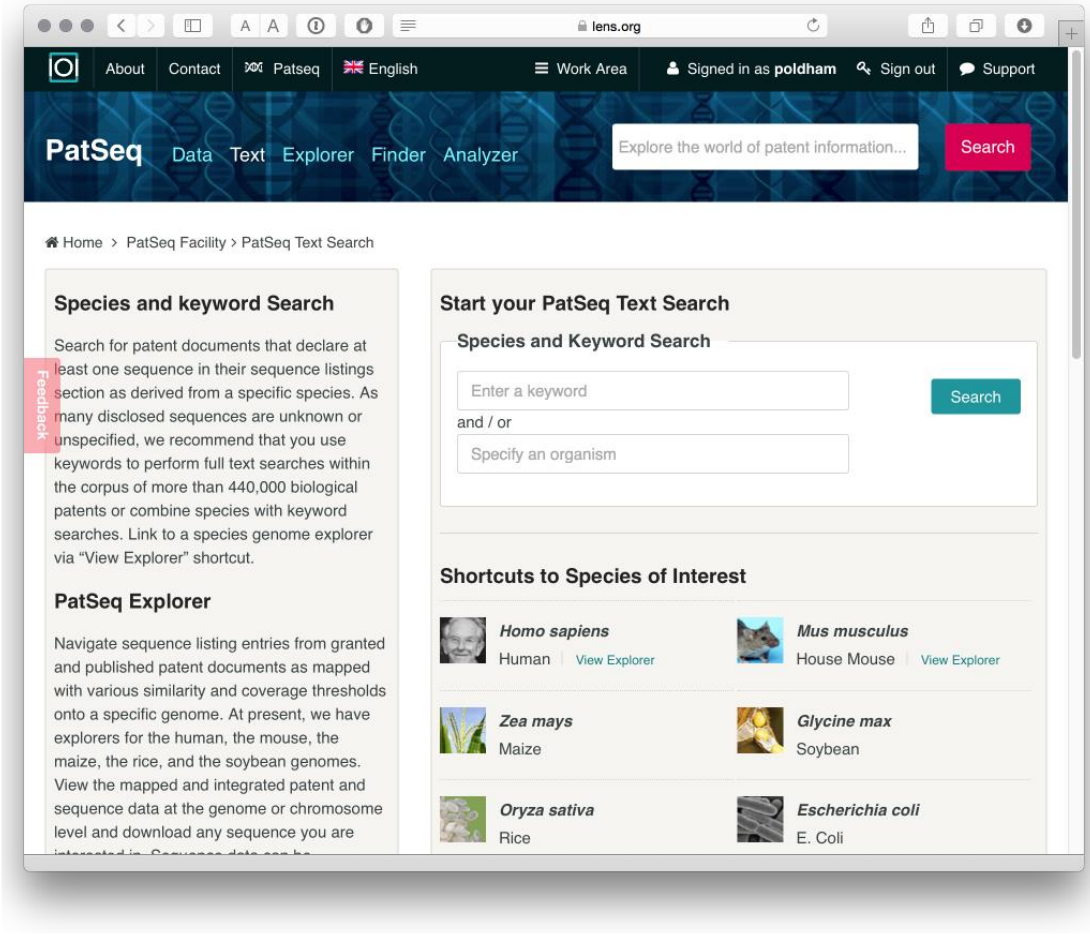
6.6 PatSeq

Un foco importante del desarrollo de Lens ha sido la secuencia de datos de ADN, incluida una serie de artículos en curso sobre la interpretación y el significado de la secuencia de datos en la actividad de patentes.



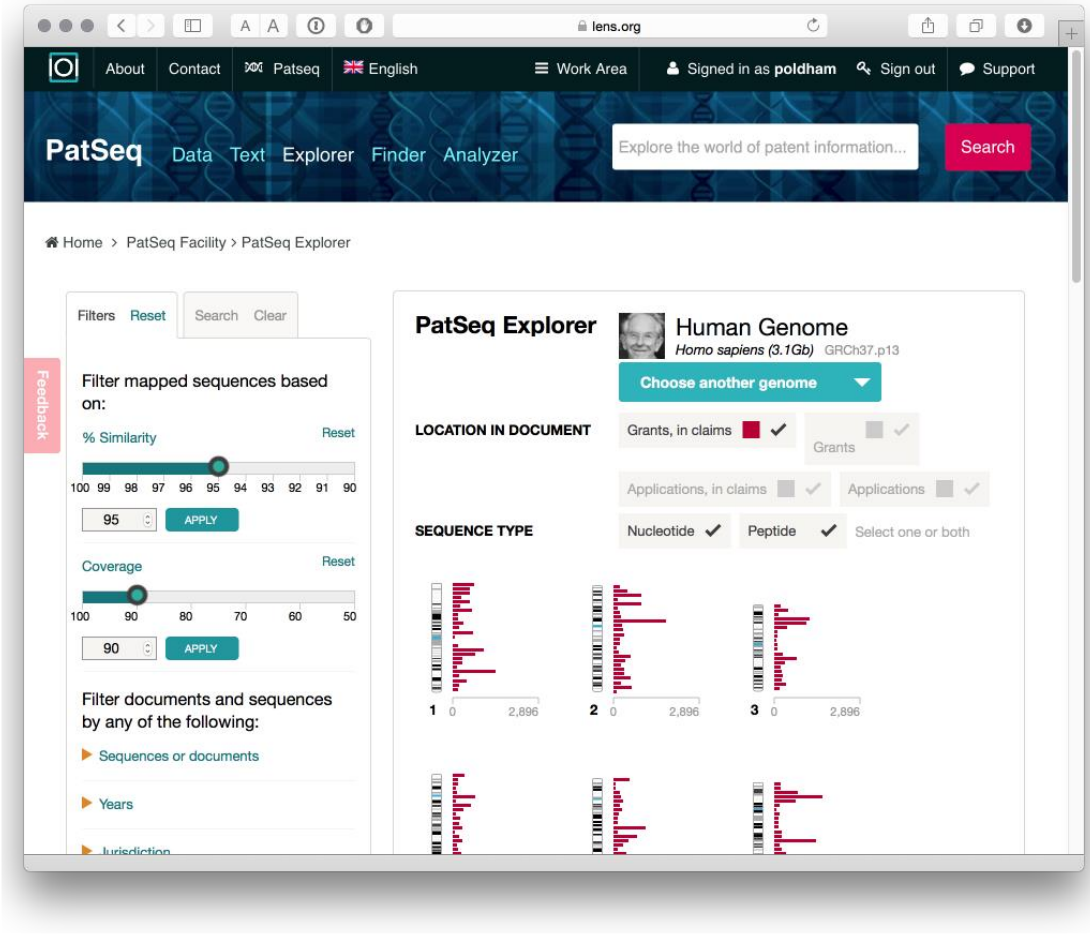
Patseq incluye una serie de herramientas.

1. Los datos de PatSeq permiten el acceso a los documentos de patentes, revelando las secuencias disponibles para descarga masiva desde un número creciente de países. Este es un sitio muy útil para obtener datos de secuencia. Tenga en cuenta que deberá solicitar acceso para descargar datos de secuencia en el área de su cuenta.
2. El buscador de especies y la búsqueda de palabras clave se centran en la búsqueda de documentos que contienen una secuencia para el nombre de una especie o término clave.



Se han generado una serie de carteras de patentes para algunas de las principales especies de plantas y animales, por ejemplo, arroz, maíz, seres humanos, pollos, etc., que pueden descargarse como colecciones.

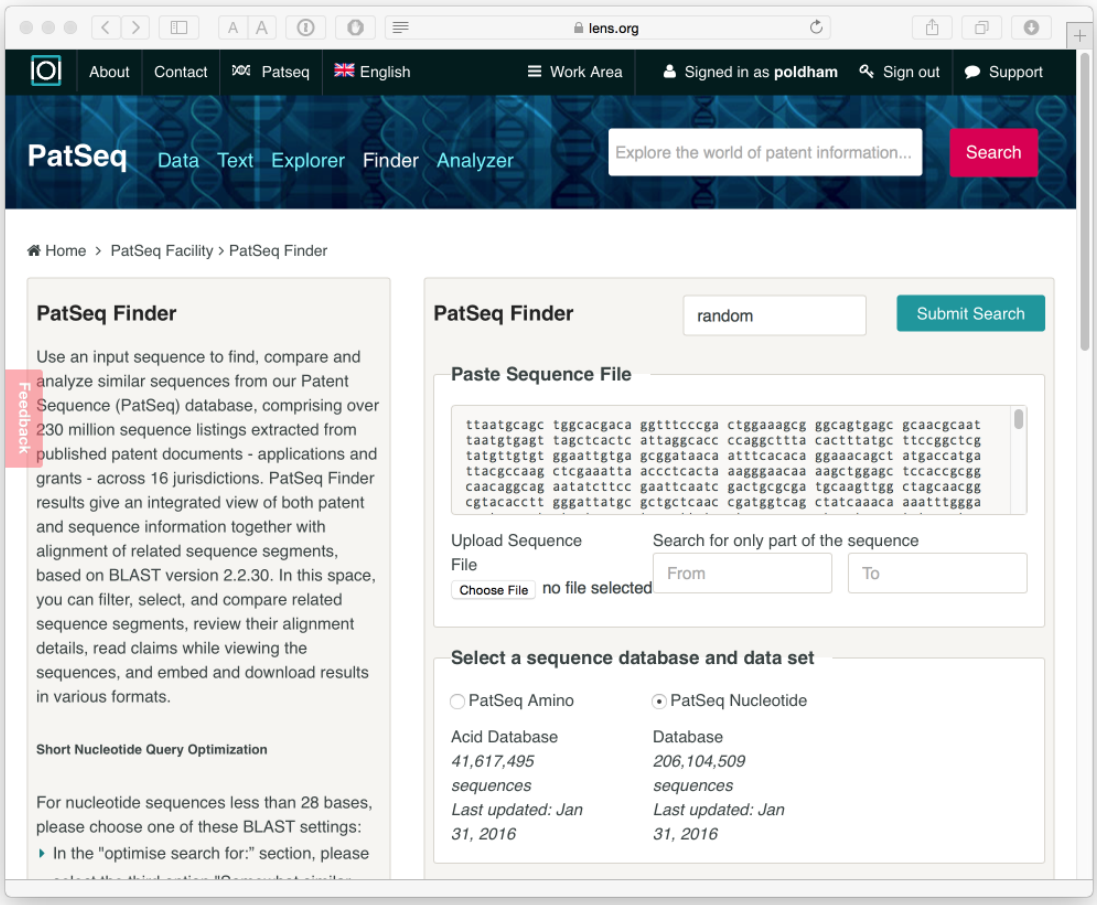
3. El Explorador de PatSeq permite la exploración de datos de secuencia para cuatro genomas (en la actualidad), especialmente el genoma humano y de ratón para animales y el genoma de soja, maíz y arroz para plantas.



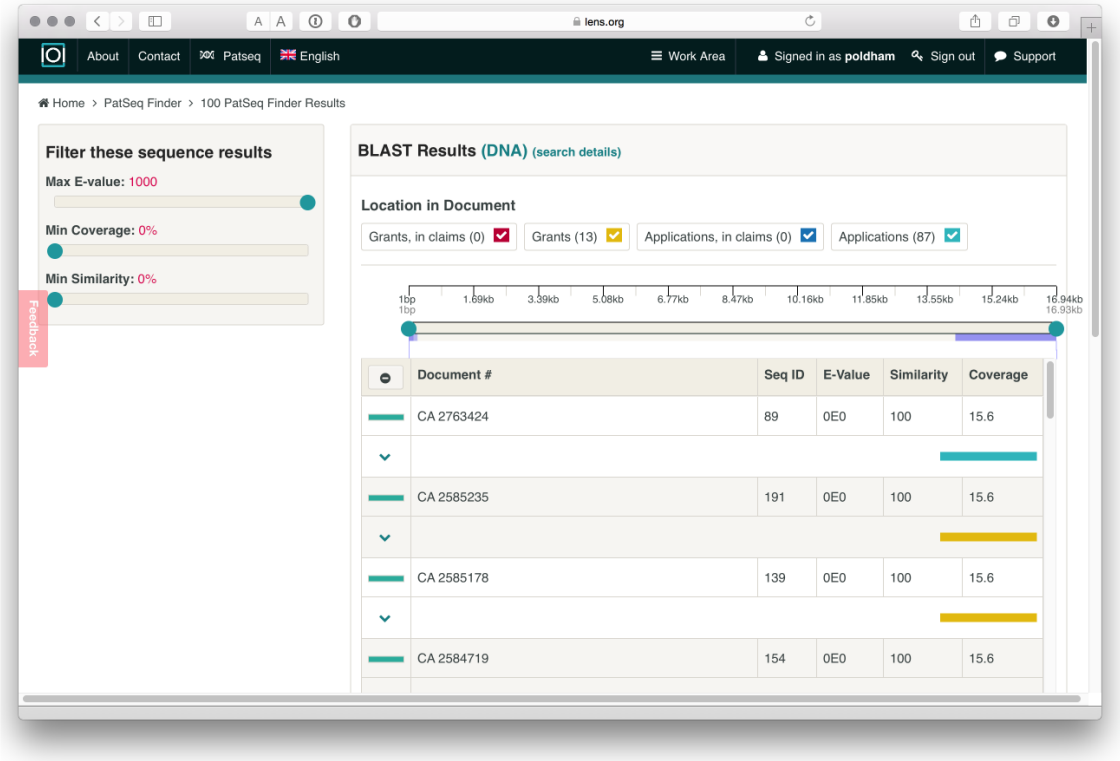
Esta es un área donde los investigadores de Cambia , la organización sin fines de lucro que está detrás de Lens, han invertido un esfuerzo considerable y vale la pena leer los artículos de investigación que aparecen en los sitios web de Cambia y Lens sobre este tema. PatSeq Analyzer está estrechamente relacionado con el Explorer y actualmente proporciona detalles sobre los genomas mencionados anteriormente con un resumen detallado de las secuencias por documento que incluye la región, secuencia, transcripción, polimorfismos de un solo nucleótido (SNP) y otorga secuencias en las reivindicaciones de patente.

4. Buscador de PatSeq

El Buscador de PatSeq permite que un usuario ingrese una secuencia de ADN o de aminoácidos en el cuadro de búsqueda y encuentre aplicaciones y subvenciones con secuencias idénticas o similares. Seleccionamos una secuencia al azar del navegador de listados de secuencias de WIPO Patentscope W016 / 026850 .



Después del procesamiento, veremos una lista de resultados que se pueden descargar en una variedad de formatos. Los resultados indican que nuestra secuencia aleatoria no aparece en las reivindicaciones de una patente concedida o una solicitud de patente, pero sí aparece en varias solicitudes y subvenciones. Se proporcionan más detalles al desplazarse sobre las entradas individuales y hay controles adicionales disponibles para la similitud y otros puntajes para refinar los resultados.



Por lo que podemos decir, mientras que los datos se pueden descargar, actualmente no es posible generar una colección de documentos a partir de los resultados del buscador de PatSeq.

6.7 Redondeo

The Lens es una base de datos de patentes muy útil que, cuando ha descubierto el significado de los íconos, es fácil de usar. La facilidad con la que se pueden compartir las colecciones y la descarga de hasta 10,000 registros es una ventaja real para el objetivo. Además, el uso de HTML5 y Highcharts hace de esta una experiencia altamente interactiva. La capacidad de usar gráficos para profundizar en los datos es muy bienvenida. El enlace al crossrefservicio para publicaciones no relacionadas con patentes es muy útil, pero sería bueno ver estos datos incluidos de alguna manera como un campo en las descargas de datos.

Con la adición de descargas de datos (en 2015), el objetivo se está convirtiendo en una plataforma muy útil para buscar, refinar, visualizar y descargar datos de patentes. Lo que quizás sería útil sería un conjunto de demostraciones o casos de uso que expliquen la forma en que se puede utilizar Lens en flujos de trabajo comunes. Por ejemplo, desarrollar y refinar una búsqueda, probar resultados, y luego recuperar citas anteriores y posteriores para el refinamiento y la

visualización son tareas bastante comunes en el análisis del panorama de patentes. Los casos de uso ayudarían a los usuarios a aprovechar al máximo lo que el objetivo tiene para ofrecer.

The Lens también destaca por su trabajo distintivo a largo plazo sobre datos de secuencia en patentes y esto será de particular interés para los investigadores que trabajan en biotecnología, particularmente en la exploración de herramientas analíticas.

Capítulo 7 Patentscope

7.1 Introducción

Patentscope es la base de datos de acceso público de la OMPI. Incluye la cobertura de las solicitudes del Tratado de Cooperación en materia de Patentes (administrado por la OMPI) y una amplia gama de otros países, incluida la Oficina Europea de Patentes, la USPTO y Japón, con un total de 51 millones de documentos de patentes, incluidos 2,8 millones de solicitudes PCT.

En este artículo cubrimos los conceptos básicos del uso de Patentscope para buscar y descargar hasta 10,000 registros. Una guía del usuario detallada proporciona más detalles sobre características específicas. Un conjunto de videos tutoriales también están disponibles. En comparación con otros servicios gratuitos, Patentscope tiene las siguientes fortalezas principales.

1. Búsqueda de texto completo en la descripción y las reclamaciones de las solicitudes PCT el día de la publicación y las solicitudes de patentes de una amplia gama de otros países, incluidos Estados Unidos, Japón, China y la Oficina Europea de Patentes, entre otros.
2. Descarga hasta 10,000 registros
3. Expande los términos de búsqueda en varios otros idiomas usando Cross Lingual Expansiono CLIR
4. Búsqueda simple, avanzada y combinada de campo
5. Accesible en múltiples idiomas y con una función de traducción de texto de la OMPI
6. Versión móvil y https: acceso
7. Descargas de listas de secuencias
8. Tecnologías verdes a través del Inventario Verde de IPC
9. Diferentes tipos de análisis gráficos de listas de resultados sobre la marcha utilizando el menú Opciones.

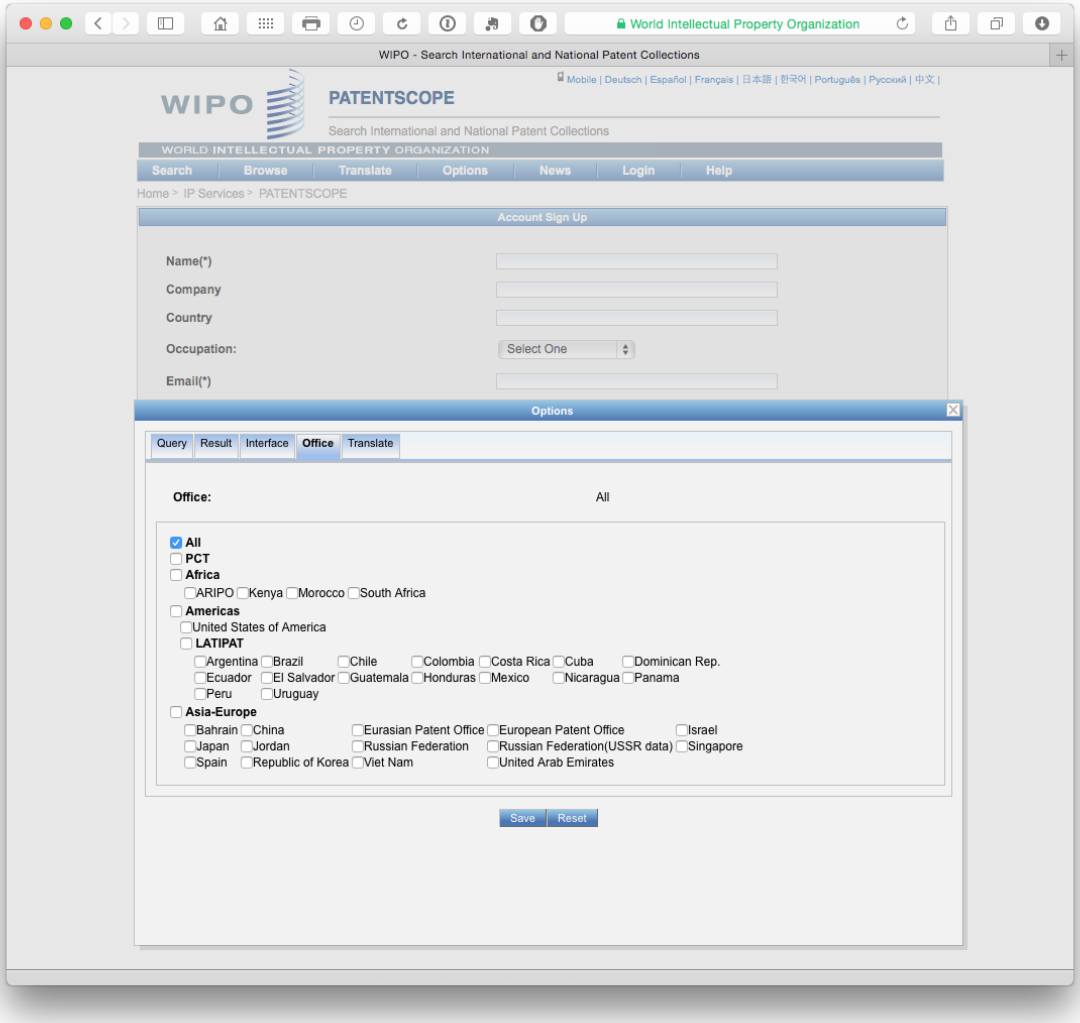
Para aprovechar al máximo Patentscope, es una buena idea consultar las dos guías detalladas y los tutoriales en video:

1. Búsqueda de Patentscope: La Guía del usuario .
2. Patentscope CLIR para la herramienta de recuperación de información en varios idiomas aquí .
3. Video tutoriales de Patentscope

Si desea descargar datos de patentes o secuencias, deberá registrarse para obtener una cuenta gratuita. Para registrarse para una cuenta gratuita vaya aquí .

7.2 Colecciones a buscar

Quizás el mejor lugar para comenzar sea con las colecciones que buscaremos. Se puede acceder a ellos en el menú de Opciones en el menú principal y luego en la pestaña de lectura de la oficina .



Aquí podemos ver que Patentscope proporciona acceso a la colección del Tratado de Cooperación en materia de Patentes, colecciones regionales como la ARIPO y la Oficina Europea de Patentes y colecciones nacionales como los Estados Unidos, Japón y otros. La capacidad de buscar y recuperar datos de la colección de LATIPAT será particularmente útil para los investigadores en América Latina y

podría vincularse al análisis utilizando la versión espacenet de LATIPAT . Si solo está interesado en colecciones particulares, este es el lugar para cambiar la configuración.

7.3 Búsqueda simple

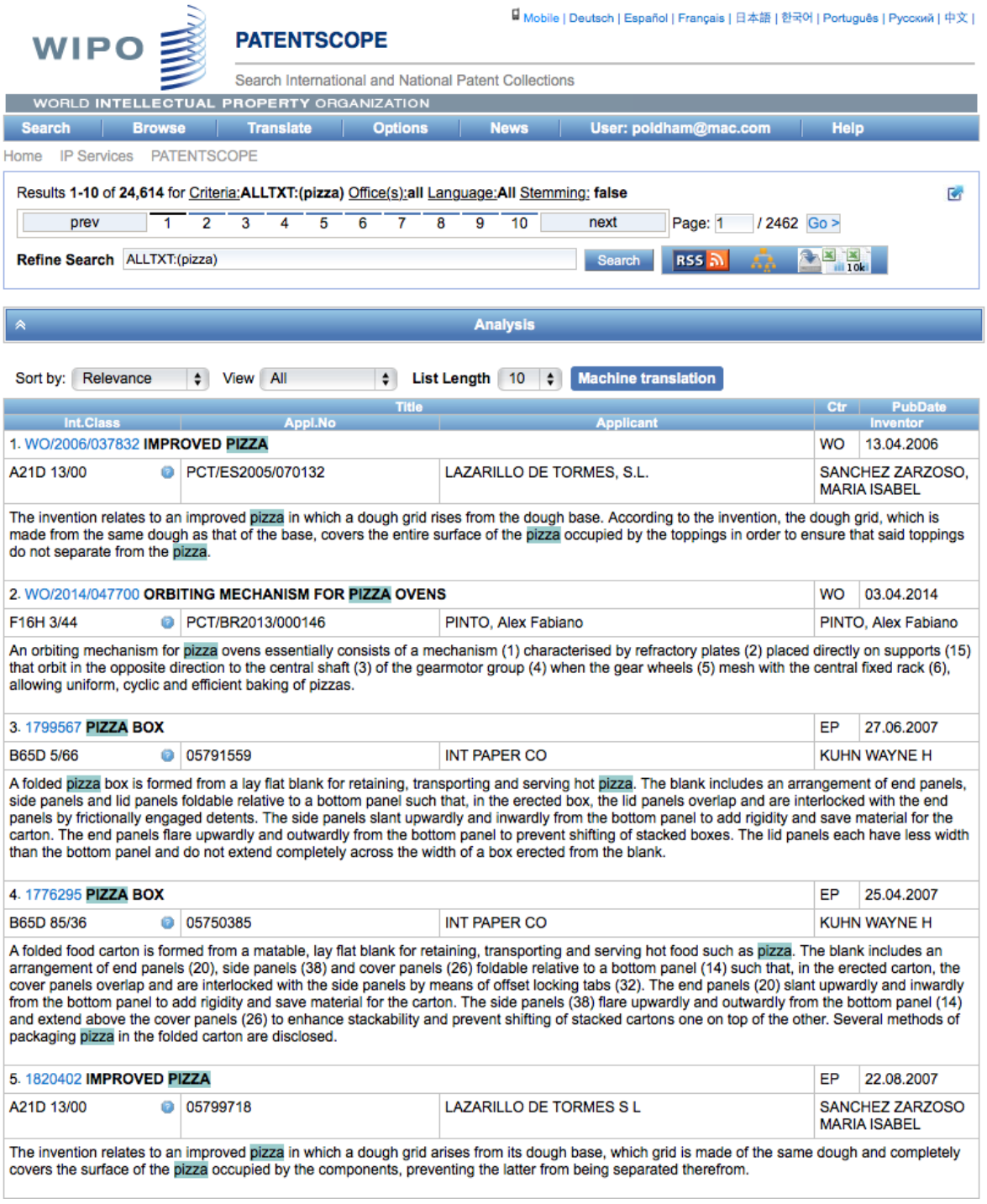
Podemos seleccionar un rango de campos diferentes para la búsqueda. En este caso, hemos seleccionado el texto completo del menú desplegable para una búsqueda simple en el término pizza.

Tenga en cuenta que Patentscope agrupa documentos para la misma aplicación en un registro o expediente y que estamos viendo el documento que es la clave para el registro. Se puede acceder a los otros documentos en el expediente para el registro haciendo clic en el número de documento y seleccionando el menú Documentos como en este ejemplo .

Para más detalles sobre el uso simple búsqueda ver el Simple Search video tutorial . Los videos también están disponibles sobre el uso de Combinaciones de campos para construir búsquedas y Búsqueda avanzada.

7.4 resultados

Cuando llegamos a la página de resultados, podemos ver que tenemos 24,614 resultados con nuestra consulta mostrando como búsqueda All TXT y todos los idiomas. Luego tenemos un botón RSS para copiar la fuente en un alimentador RSS para actualizaciones.



WIPO PATENTSCOPE

Mobile | Deutsch | Español | Français | 日本語 | 한국어 | Português | Русский | 中文 |

Search International and National Patent Collections

WORLD INTELLECTUAL PROPERTY ORGANIZATION

Search | Browse | Translate | Options | News | User: poldham@mac.com | Help

Home IP Services PATENTSCOPE

Results 1-10 of 24,614 for Criteria:ALLTXT:(pizza) Office(s):all Language:All Stemming: false

prev 1 2 3 4 5 6 7 8 9 10 next Page: 1 / 2462 Go >

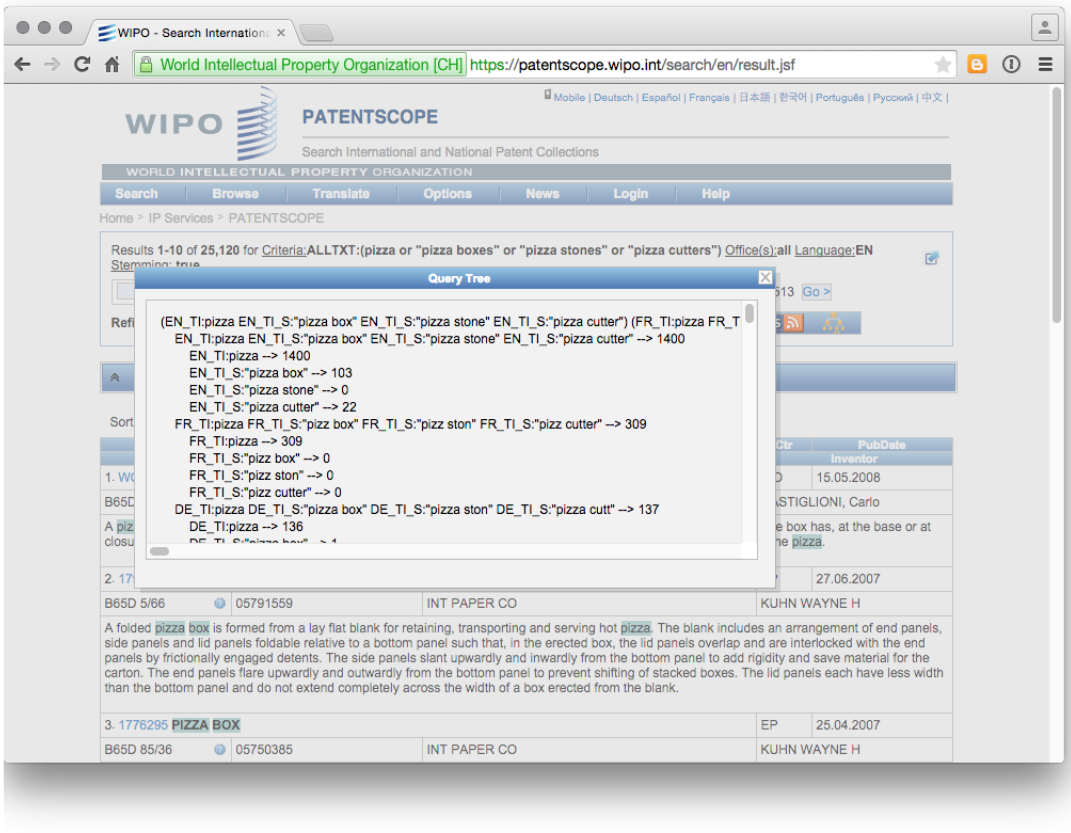
Refine Search ALLTXT:(pizza) Search RSS 10k

Analysis

Sort by: Relevance View All List Length 10 Machine translation

Int.Class	Appl.No	Applicant	Ctr / Inventor	PubDate
1. WO/2006/037832 **IMPROVED PIZZA**			WO	13.04.2006
A21D 13/00	PCT/ES2005/070132	LAZARILLO DE TORMES, S.L.	SANCHEZ ZARZOSO, MARIA ISABEL	
The invention relates to an improved pizza in which a dough grid rises from the dough base. According to the invention, the dough grid, which is made from the same dough as that of the base, covers the entire surface of the pizza occupied by the toppings in order to ensure that said toppings do not separate from the pizza.				
2. WO/2014/047700 **ORBITING MECHANISM FOR PIZZA OVENS**			WO	03.04.2014
F16H 3/44	PCT/BR2013/000146	PINTO, Alex Fabiano	PINTO, Alex Fabiano	
An orbiting mechanism for pizza ovens essentially consists of a mechanism (1) characterised by refractory plates (2) placed directly on supports (15) that orbit in the opposite direction to the central shaft (3) of the gearmotor group (4) when the gear wheels (5) mesh with the central fixed rack (6), allowing uniform, cyclic and efficient baking of pizzas.				
3. 1799567 **PIZZA BOX**			EP	27.06.2007
B65D 5/66	05791559	INT PAPER CO	KUHN WAYNE H	
A folded pizza box is formed from a lay flat blank for retaining, transporting and serving hot pizza. The blank includes an arrangement of end panels, side panels and lid panels foldable relative to a bottom panel such that, in the erected box, the lid panels overlap and are interlocked with the end panels by frictionally engaged detents. The side panels slant upwardly and inwardly from the bottom panel to add rigidity and save material for the carton. The end panels flare upwardly and outwardly from the bottom panel to prevent shifting of stacked boxes. The lid panels each have less width than the bottom panel and do not extend completely across the width of a box erected from the blank.				
4. 1776295 **PIZZA BOX**			EP	25.04.2007
B65D 85/36	05750385	INT PAPER CO	KUHN WAYNE H	
A folded food carton is formed from a matable, lay flat blank for retaining, transporting and serving hot food such as pizza. The blank includes an arrangement of end panels (20), side panels (38) and cover panels (26) foldable relative to a bottom panel (14) such that, in the erected carton, the cover panels overlap and are interlocked with the side panels by means of offset locking tabs (32). The end panels (20) slant upwardly and inwardly from the bottom panel to add rigidity and save material for the carton. The side panels (38) flare upwardly and outwardly from the bottom panel (14) and extend above the cover panels (26) to enhance stackability and prevent shifting of stacked cartons one on top of the other. Several methods of packaging pizza in the folded carton are disclosed.				
5. 1820402 **IMPROVED PIZZA**			EP	22.08.2007
A21D 13/00	05799718	LAZARILLO DE TORMES S L	SANCHEZ ZARZOSO MARIA ISABEL	
The invention relates to an improved pizza in which a dough grid arises from its dough base, which grid is made of the same dough and completely covers the surface of the pizza occupied by the components, preventing the latter from being separated therefrom.				

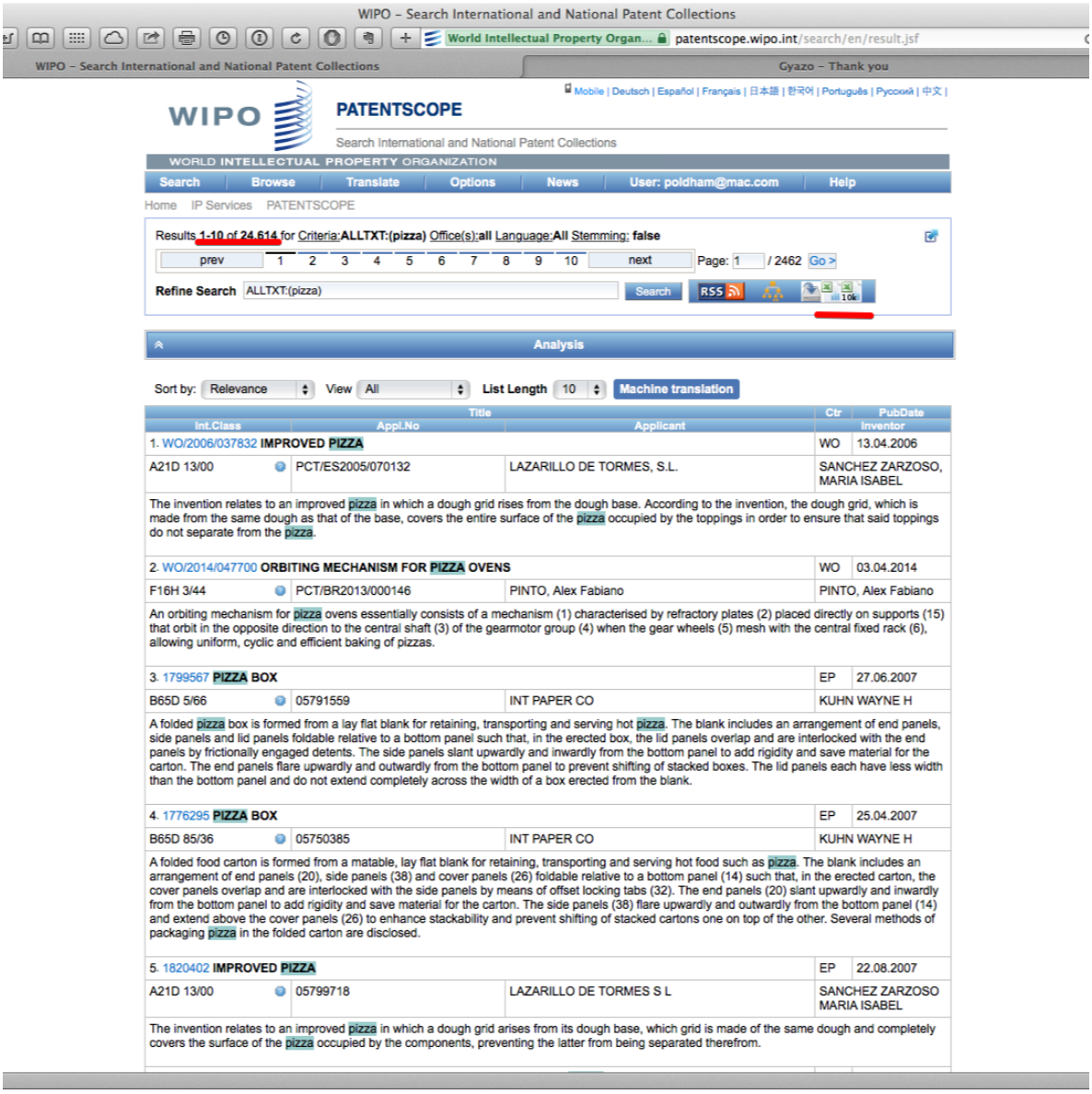
También hay un botón de árbol de consulta que muestra los resultados por idioma y términos en las secciones relevantes del documento. Podemos ver un ejemplo de esto para una consulta más compleja a continuación.



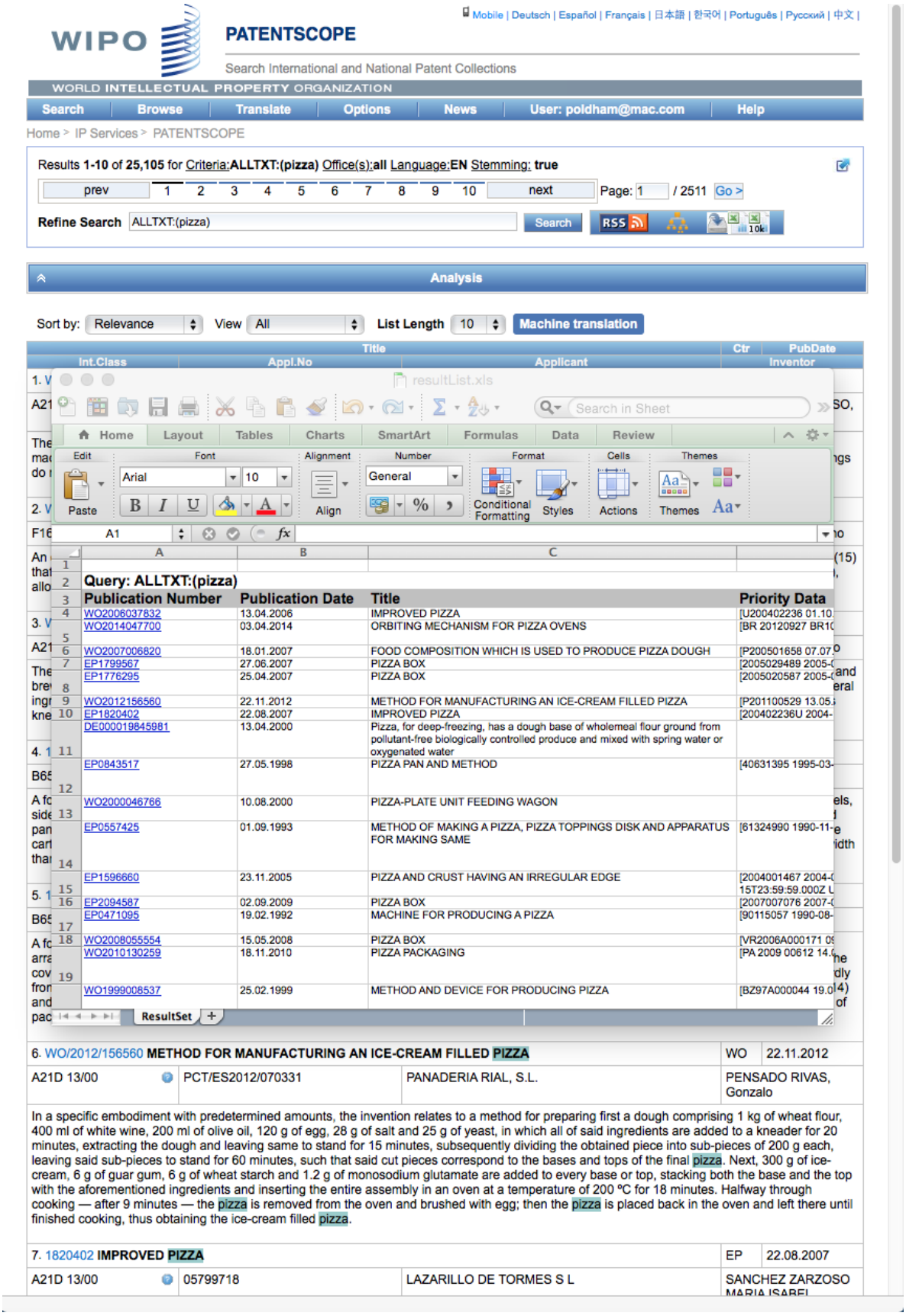
Un video tutorial también está disponible para la lista de resultados de búsqueda

7.5 Descargando Resultados

Los dos iconos de Excel al final del menú permiten al usuario descargar la lista corta (primer icono) o la segunda lista como un archivo .xls. Para ver estos íconos, debe iniciar sesión con una cuenta de usuario o no se mostrarán.



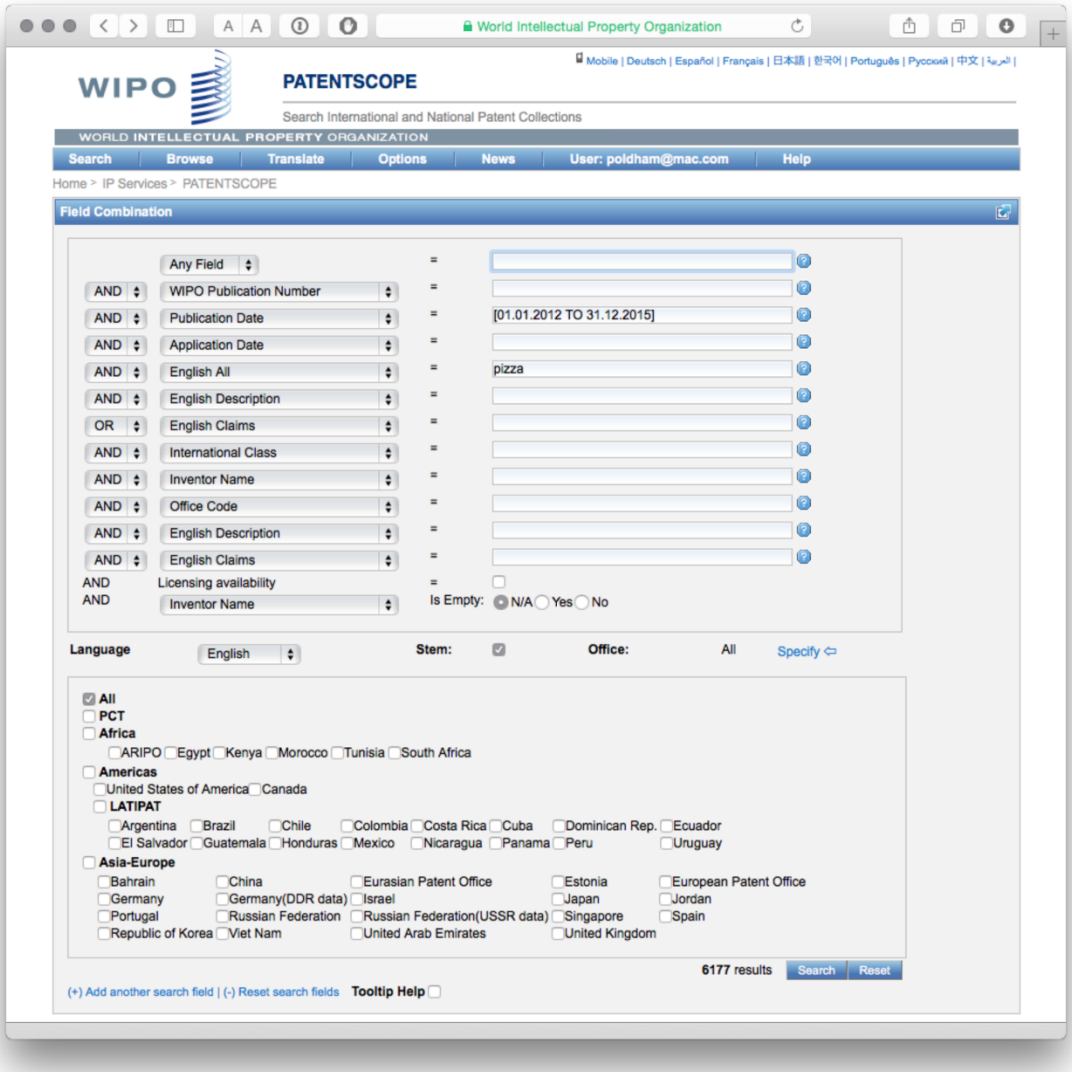
Cuando descarguemos estos resultados, recibiremos una hoja .xls con hasta 10,000 entradas con un par de filas de encabezado que muestran la consulta. Tenga en cuenta que cada registro en la hoja de Excel está hipervinculado al registro correspondiente en Patentscope.


WIPO
PATENTSCOPE
Search International and National Patent Collections
WORLD INTELLECTUAL PROPERTY ORGANIZATION
Search | Browse | Translate | Options | News | User: poldham@mac.com | Help
Results 1-10 of 25,105 for Criteria:ALLTXT:(pizza) Office(s):all Language:EN Stemming: true
Refine Search ALLTXT:(pizza)
Analysis
Query: ALLTXT:(pizza)
Publication Number | Publication Date | Title | Priority Data
WO2006037832 | 13.04.2006 | IMPROVED PIZZA
WO2014047700 | 03.04.2014 | ORBITING MECHANISM FOR PIZZA OVENS
WO2007006820 | 18.01.2007 | FOOD COMPOSITION WHICH IS USED TO PRODUCE PIZZA DOUGH
EP1799567 | 27.06.2007 | PIZZA BOX
EP1776295 | 25.04.2007 | PIZZA BOX
WO2012156560 | 22.11.2012 | METHOD FOR MANUFACTURING AN ICE-CREAM FILLED PIZZA
EP1820402 | 22.08.2007 | IMPROVED PIZZA
DE000019845981 | 13.04.2000 | Pizza, for deep-freezing, has a dough base of wholemeal flour ground from pollutant-free biologically controlled produce and mixed with spring water or oxygenated water
EP0843517 | 27.05.1998 | PIZZA PAN AND METHOD
WO2000046766 | 10.08.2000 | PIZZA-PLATE UNIT FEEDING WAGON
EP0557425 | 01.09.1993 | METHOD OF MAKING A PIZZA, PIZZA TOPPINGS DISK AND APPARATUS FOR MAKING SAME
EP1596660 | 23.11.2005 | PIZZA AND CRUST HAVING AN IRREGULAR EDGE
EP2094587 | 02.09.2009 | PIZZA BOX
EP0471095 | 19.02.1992 | MACHINE FOR PRODUCING A PIZZA
WO2008055554 | 15.05.2008 | PIZZA BOX
WO2010130259 | 18.11.2010 | PIZZA PACKAGING
WO1999008537 | 25.02.1999 | METHOD AND DEVICE FOR PRODUCING PIZZA
6. WO/2012/156560 METHOD FOR MANUFACTURING AN ICE-CREAM FILLED PIZZA
WO 22.11.2012
A21D 13/00 PCT/ES2012/070331 PANADERIA RIAL, S.L. PENSADO RIVAS, Gonzalo
In a specific embodiment with predetermined amounts, the invention relates to a method for preparing first a dough comprising 1 kg of wheat flour, 400 ml of white wine, 200 ml of olive oil, 120 g of egg, 28 g of salt and 25 g of yeast, in which all of said ingredients are added to a kneader for 20 minutes, extracting the dough and leaving same to stand for 15 minutes, subsequently dividing the obtained piece into sub-pieces of 200 g each, leaving said sub-pieces to stand for 60 minutes, such that said cut pieces correspond to the bases and tops of the final pizza. Next, 300 g of ice-cream, 6 g of guar gum, 6 g of wheat starch and 1.2 g of monosodium glutamate are added to every base or top, stacking both the base and the top with the aforementioned ingredients and inserting the entire assembly in an oven at a temperature of 200 ºC for 18 minutes. Halfway through cooking — after 9 minutes — the pizza is removed from the oven and brushed with egg; then the pizza is placed back in the oven and left there until finished cooking, thus obtaining the ice-cream filled pizza.
7. 1820402 IMPROVED PIZZA
EP 22.08.2007
A21D 13/00 05799718 LAZARILLO DE TORMES S L SANCHEZ ZARZOSO

Entraremos en el uso de estos datos, incluso con Tableau Public y otras herramientas, con cierta profundidad y hay algunas cosas a tener en cuenta aquí. La primera es que el número de publicación con hipervínculo no posee un código de tipo (A1, B1, etc.). Esto solo importa en el sentido de que el número recuperará varios documentos en otras bases de datos vinculadas al número de Patentscope. Un segundo punto para destacar es que los datos de Patentscope son raw en el sentido de que son datos, ya que provienen de los proveedores de datos y no se procesan. Eso significa que puede haber problemas de codificación a los que volveremos más adelante en las discusiones sobre la limpieza de datos.

Lo que es muy útil acerca de Patentscope es que podemos obtener un volumen de datos bastante significativo sobre un tema de interés. Si bien este artículo simplemente ha descargado los primeros 10,000 resultados, para obtener el conjunto completo de resultados sería bastante fácil limitar los datos por año y descargar los datos como una serie de conjuntos que se pueden combinar más tarde (por ejemplo, tres conjuntos).

Para hacer esto necesitamos visitar la página de combinación de campos. Aquí comenzaremos poniendo nuestra consulta en inglés Todos para obtener el número total de resultados. Luego restringiremos los datos por el campo de datos de publicación utilizando []y un período entre fechas (como DD.MM.YYYY). A continuación, se muestra un ejemplo.



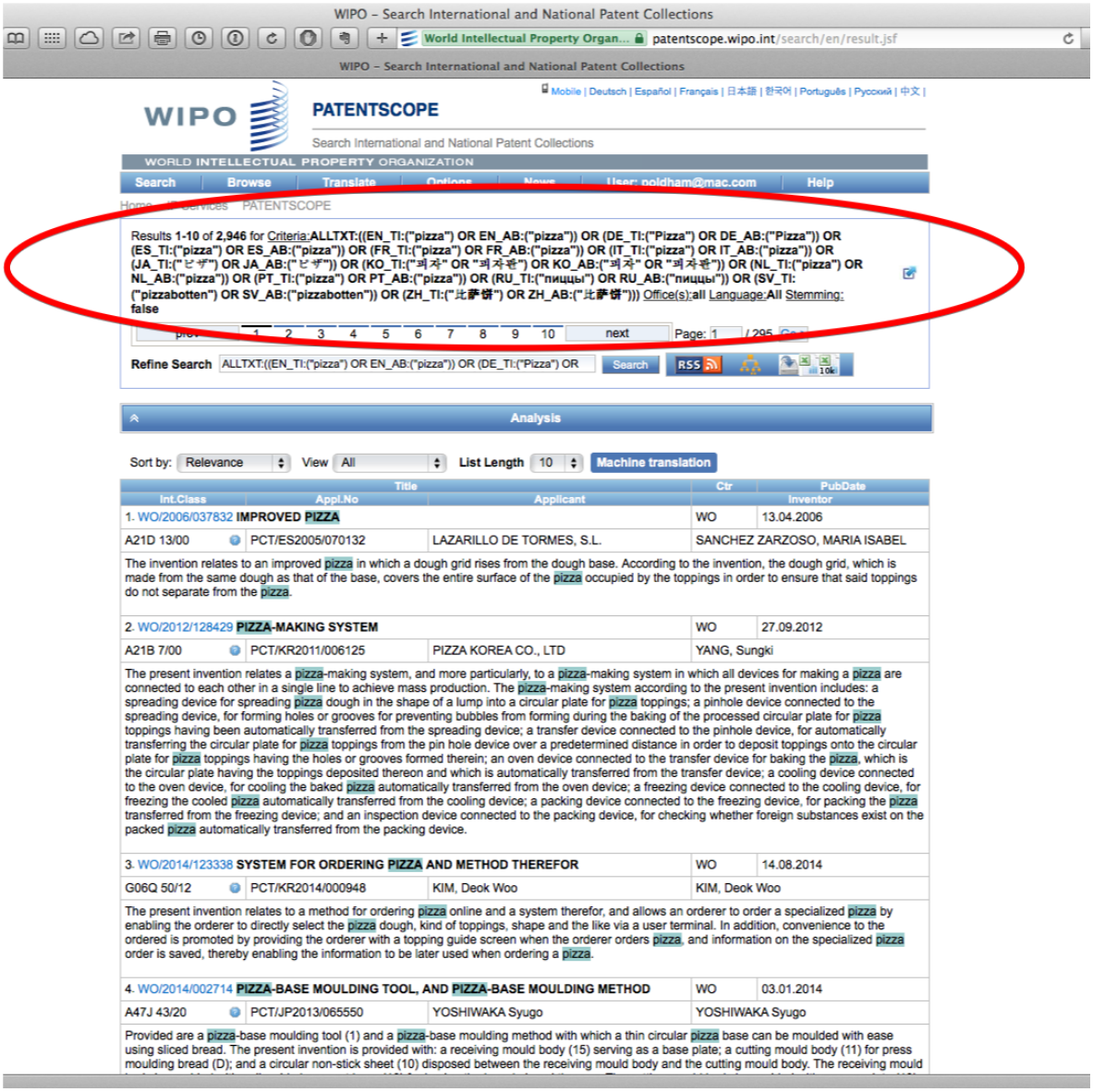
Esto mostrará de manera útil los resultados totales de la consulta (aunque puede llevar algo de tiempo) y podemos ejecutar y luego descargar los resultados para cada segmento limitado del año.

Cuando trabaje con descargas múltiples, es una buena idea anotar el número total de resultados y luego los resultados para cada segmento de fecha limitada para asegurar que los datos se sumen a lo que usted esperaría. También es posible que se necesite algo de experimentación con la configuración de campo usando los operadores booleanos AND / OR.

7.6 Búsqueda lingual cruzada

Un desafío en la búsqueda de patentes es el uso de diferentes expresiones en diferentes idiomas para la misma consulta. Patentscope presenta una solución muy útil para esto a través de la búsqueda en varios idiomas. En el menú desplegable, seleccione Cross Lingual Expansion, luego ingrese los términos de búsqueda y

presione ir. La herramienta ahora generará términos de búsqueda en múltiples idiomas.



Para ir más lejos con esta herramienta, use la configuración del control deslizante (precisión vs. recuperación). Por ejemplo, si insertáramos el término de búsqueda "biología sintética" y moviéramos la recuperación al nivel superior (4), generaríamos la siguiente consulta.

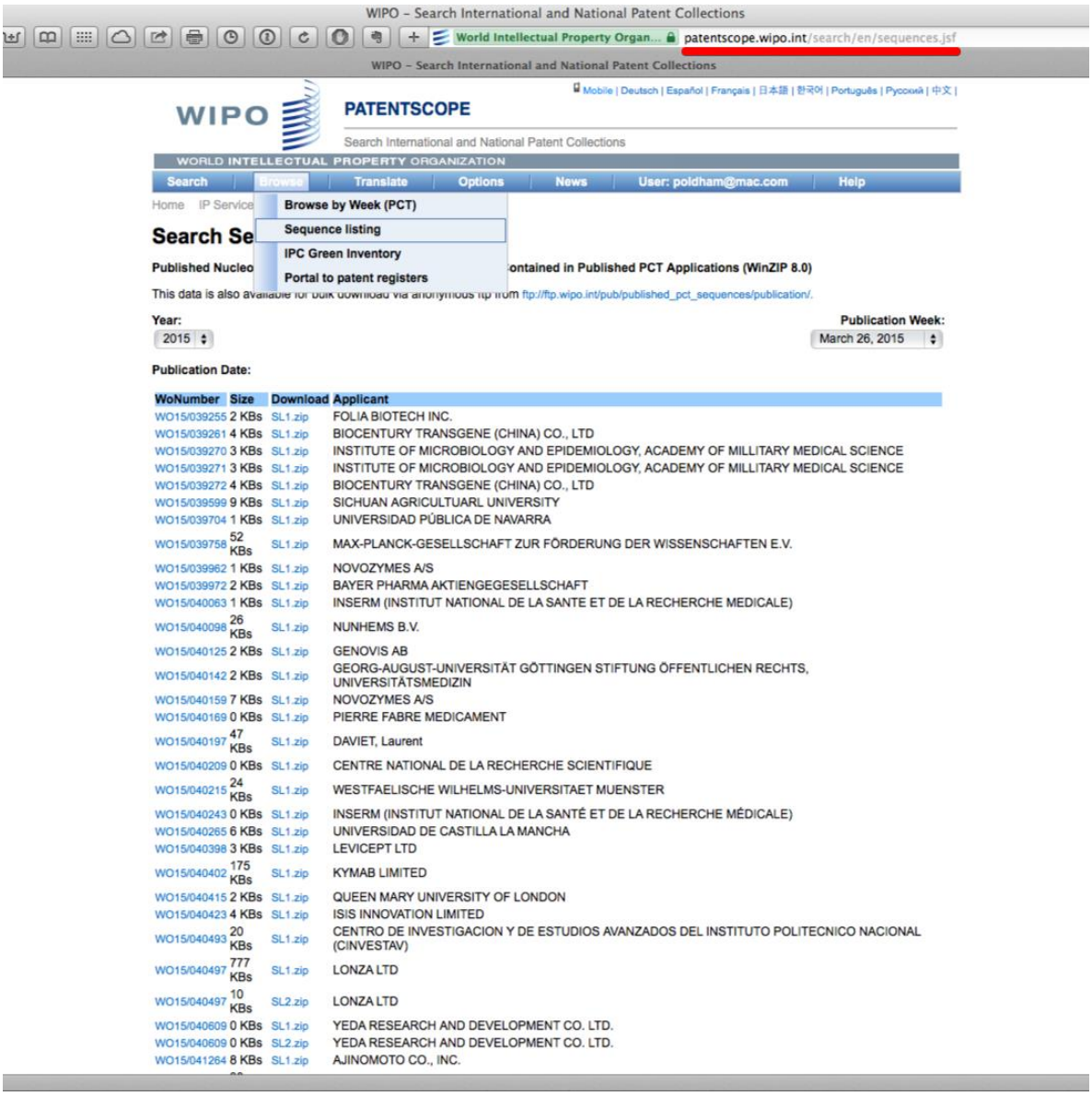
"FP:((EN_TI:("synthetic biology" OR "biologic synthetic") OR EN_AB:("synthetic biology" OR "biologic synthetic")) OR (DE_TI:("synthetische Biologie" OR "synthetischen biologischen" OR "biologische synthetische" OR "Biologische synthetische") OR DE_AB:("synthetische Biologie" OR

"synthetischen biologischen" OR "biologische synthetische" OR "Biologische synthetische")) OR (ES_TI:("biológicas sintéticas") OR ES_AB:("biológicas sintéticas")) OR (FR_TI:("biologie synthétique" OR "biologie synthéthique") OR FR_AB:("biologie synthétique" OR "biologie synthéthique")) OR (JA_TI:("生物合成" OR "合成生体" OR "の生物学的合成") OR JA_AB:("生物合成" OR "合成生体" OR "の生物学的合成")) OR (ZH_TI:("合成生物") OR ZH_AB:("合成生物")))"

Si se selecciona el modo supervisado en el Expansion modemenú desplegable, es posible seleccionar áreas de tecnología para la generación de terminología. Si bien no hemos analizado esto en detalle, esto podría ser muy útil para la generación de consultas específicas del dominio. En definitiva, esta es una de las herramientas más originales y poderosas que ofrece Patentscope. Una .pdfguía detallada para usar CLIR está disponible aquí .

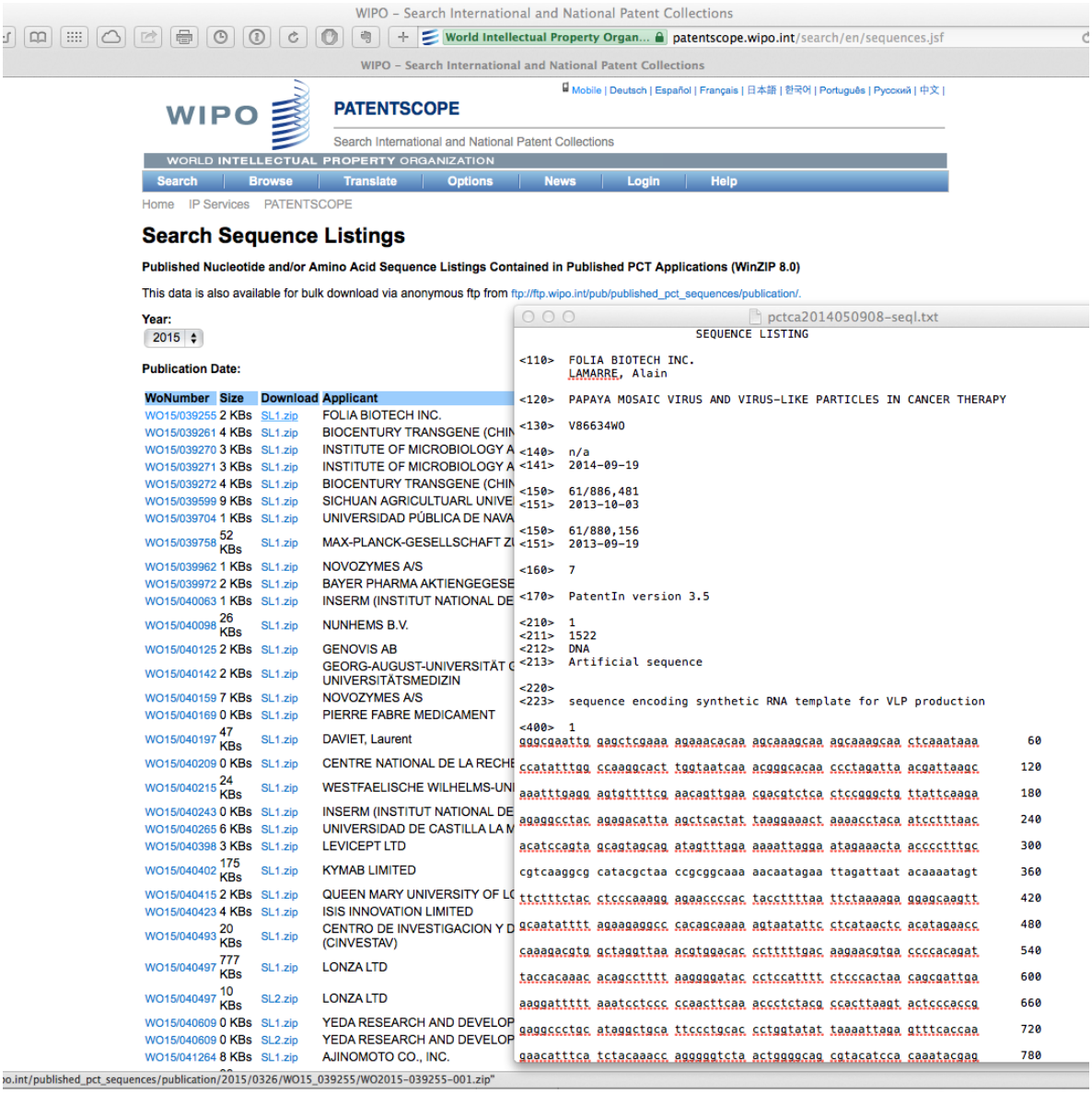
7.7 Datos de secuencia

Una tercera característica importante de Patentscope es el acceso a las listas de secuencias de aminoácidos y ADN archivadas con las solicitudes PCT. Estos datos pueden consultarse y descargarse para los registros individuales aquí .

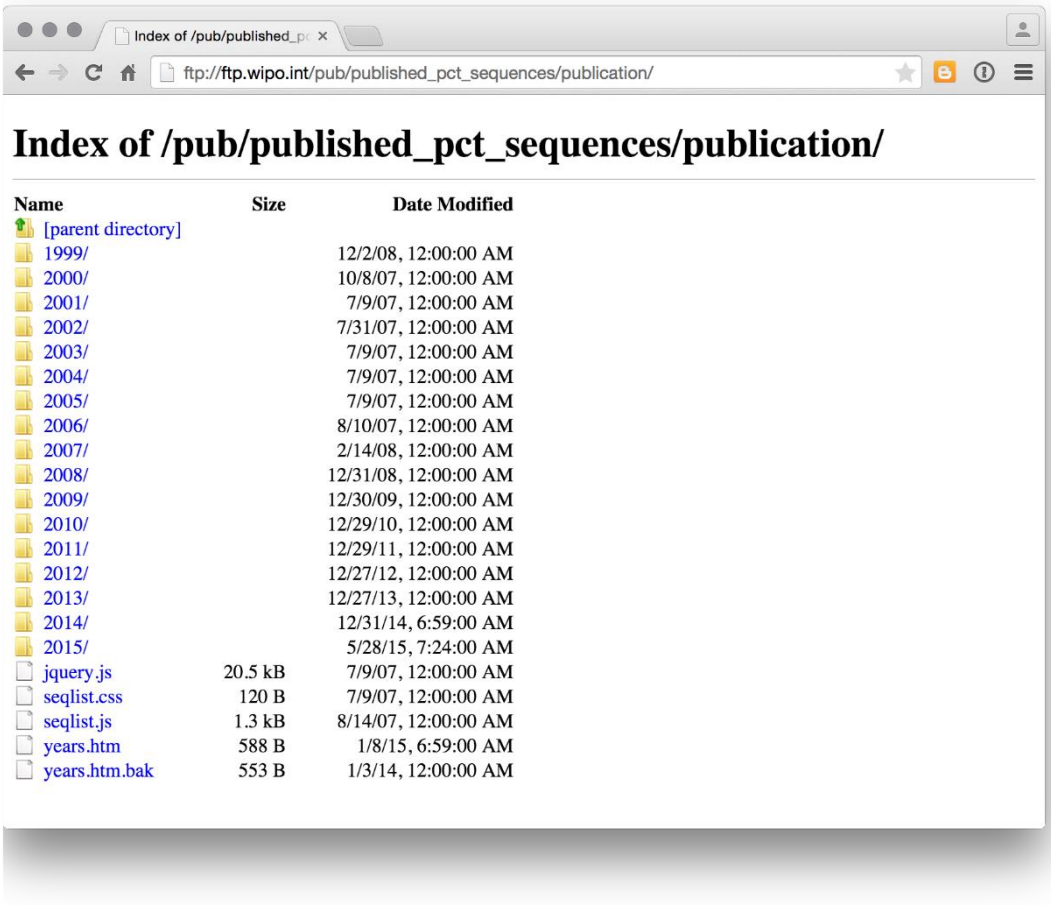


WoNumber	Size	Download	Applicant
WO15/039255	2 KBs	SL1.zip	FOLIA BIOTECH INC.
WO15/039261	4 KBs	SL1.zip	BIOCENTURY TRANSGENE (CHINA) CO., LTD
WO15/039270	3 KBs	SL1.zip	INSTITUTE OF MICROBIOLOGY AND EPIDEMIOLOGY, ACADEMY OF MILLITARY MEDICAL SCIENCE
WO15/039271	3 KBs	SL1.zip	INSTITUTE OF MICROBIOLOGY AND EPIDEMIOLOGY, ACADEMY OF MILLITARY MEDICAL SCIENCE
WO15/039272	4 KBs	SL1.zip	BIOCENTURY TRANSGENE (CHINA) CO., LTD
WO15/039599	9 KBs	SL1.zip	SICHUAN AGRICULTUARL UNIVERSITY
WO15/039704	1 KBs	SL1.zip	UNIVERSIDAD PÚBLICA DE NAVARRA
WO15/039758	52 KBs	SL1.zip	MAX-PLANCK-GESELLSCHAFT ZUR FÖRDERUNG DER WISSENSCHAFTEN E.V.
WO15/039962	1 KBs	SL1.zip	NOVOZYMES A/S
WO15/039972	2 KBs	SL1.zip	BAYER PHARMA AKTIENGEGESELLSCHAFT
WO15/040063	1 KBs	SL1.zip	INSERM (INSTITUT NATIONAL DE LA SANTE ET DE LA RECHERCHE MEDICALE)
WO15/040098	26 KBs	SL1.zip	NUNHEMS B.V.
WO15/040125	2 KBs	SL1.zip	GENOVIS AB
WO15/040142	2 KBs	SL1.zip	GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN STIFTUNG ÖFFENTLICHEN RECHTS, UNIVERSITÄTSMEDIZIN
WO15/040159	7 KBs	SL1.zip	NOVOZYMES A/S
WO15/040169	0 KBs	SL1.zip	PIERRE FABRE MEDICAMENT
WO15/040197	47 KBs	SL1.zip	DAVIET, Laurent
WO15/040209	0 KBs	SL1.zip	CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE
WO15/040215	24 KBs	SL1.zip	WESTFAELISCHE WILHELMS-UNIVERSITAET MUENSTER
WO15/040243	0 KBs	SL1.zip	INSERM (INSTITUT NATIONAL DE LA SANTÉ ET DE LA RECHERCHE MÉDICALE)
WO15/040265	6 KBs	SL1.zip	UNIVERSIDAD DE CASTILLA LA MANCHA
WO15/040398	3 KBs	SL1.zip	LEVICEPT LTD
WO15/040402	175 KBs	SL1.zip	KYMAB LIMITED
WO15/040415	2 KBs	SL1.zip	QUEEN MARY UNIVERSITY OF LONDON
WO15/040423	4 KBs	SL1.zip	ISIS INNOVATION LIMITED
WO15/040493	20 KBs	SL1.zip	CENTRO DE INVESTIGACION Y DE ESTUDIOS AVANZADOS DEL INSTITUTO POLITECNICO NACIONAL (CINVESTAV)
WO15/040497	777 KBs	SL1.zip	LONZA LTD
WO15/040497	10 KBs	SL2.zip	LONZA LTD
WO15/040609	0 KBs	SL1.zip	YEDA RESEARCH AND DEVELOPMENT CO. LTD.
WO15/040609	0 KBs	SL2.zip	YEDA RESEARCH AND DEVELOPMENT CO. LTD.
WO15/041264	8 KBs	SL1.zip	AJINOMOTO CO., INC.

Un registro de muestra de las listas se puede ver a continuación como un archivo de texto plano. Tenga en cuenta que pueden surgir algunos problemas al conciliar el archivo de texto plano con el número de publicación de la OMPI (WO, etc.) y esto merece una cuidadosa atención si se utilizan estos datos.



Los titulares de cuentas registrados también pueden usar el ftp anonymous downloadservicio desde la misma página. Esto proporciona acceso a los datos de la secuencia por año, como se puede ver a continuación.



Si usa el servicio ftp anónimo, tenga en cuenta que los datos recientes se miden en gigabytes, así que no intente descargarlos a través de una conexión WIFI débil, una conexión cerrada o a su teléfono (!). Sin embargo, la accesibilidad abierta de estos datos es importante. Para otras fuentes de datos de secuencia, puede estar interesado en los recursos del Instituto Europeo de Bioinformática aquí y en los EE. UU. Por el número de documento aquí y hasta marzo de 2015 en la Base de datos de patentes de ADN aquí . También es importante la Patseq herramienta Lens aquí .

7.8 Redondeo

WIPO Patentscope es una herramienta poderosa para obtener acceso a una cantidad significativa de datos de patentes sobre un tema de interés. La capacidad de descargar 10.000 o más registros a la vez no puede ser superada por otras herramientas gratuitas. La Cross Lingual Searching herramienta parece ser única y valiosa. Es probable que el acceso gratuito a la descarga masiva de datos de secuencias mantenga a los bioinformáticos felices durante bastante tiempo.

Una forma de pensar sobre el papel de Patentscope en el análisis de patentes es como un recurso que se puede combinar con otras herramientas de datos. Por ejemplo, si quisiéramos obtener los resúmenes, descripciones o reclamos de documentos PCT en Patentscope, entonces podríamos usar los números de Patenscope para recuperar datos de EPO Open Patent Services o Google Patents usando R o Python u otras herramientas. Es decir, en este caso, Patentscope supera las limitaciones de los resultados de búsqueda de otras herramientas, pero permite el uso específico de otras herramientas para recuperar más información. La Cross Lingual Searchingherramienta también podría ser particularmente útil para tratar de identificar y, posteriormente, adquirir documentos de patentes de otras jurisdicciones donde una compañía u organización puede estar buscando operar o expandir el análisis del panorama de patentes en jurisdicciones con alfabetos no romanos.

Las principales dificultades que surgen del uso de Patentscope pueden deberse a ruidos ocasionales en los datos. Patenscope no limpia los datos proporcionados por las colecciones individuales, con la excepción de la comprobación de errores tipológicos en los números de prioridad y los códigos IPC. Además, todo el texto se transforma en UTF-8. Sin embargo, como es común cuando se trata de diversas fuentes de datos, los resultados no siempre son perfectos. Además, debido a que los datos de Patentscope se obtienen de una amplia gama de idiomas, los usuarios pueden necesitar actualizar sus bibliotecas de fuentes si aparecen grandes cantidades de caracteres inusuales en los datos (como la instalación del paquete de idiomas asiáticos para Windows). En la práctica, como es común con la mayoría de las fuentes de datos de patentes, esto puede significar que se requiere un tiempo significativo para limpiar los datos. Habiendo dicho esto, ninguna otra herramienta de base de datos gratuita nos permite descargar tantos datos en forma de tabla para su análisis. Como veremos, es posible hacer mucho con los datos de Patentscope.

Capítulo 8 Abrir Refinar

La limpieza de los datos de patentes es una de las tareas más difíciles y que llevan más tiempo involucradas en el análisis de patentes. En este capítulo vamos a cubrir.

1. Limpieza básica de datos utilizando Open Refine
2. Separar un conjunto de datos de patentes en los nombres de los solicitantes y limpiar los nombres.
3. Exportación de un conjunto de datos desde Open Refine en diferentes etapas del proceso de limpieza.

Open Refine es una herramienta de código abierto para trabajar con todo tipo de datos desordenados. Comenzó su vida como Google Refine, pero desde entonces ha migrado a Open Refine. Es un programa que se ejecuta en un navegador en su computadora, pero no requiere conexión a Internet. Es una herramienta clave en el kit de herramientas de análisis de patentes de código abierto e incluye extensiones y el uso de código personalizado para tareas personalizadas particulares. En este artículo cubriremos algunos de los conceptos básicos que son más relevantes para el análisis de patentes y luego pasaremos a un trabajo más detallado para limpiar los nombres de los solicitantes de patentes.

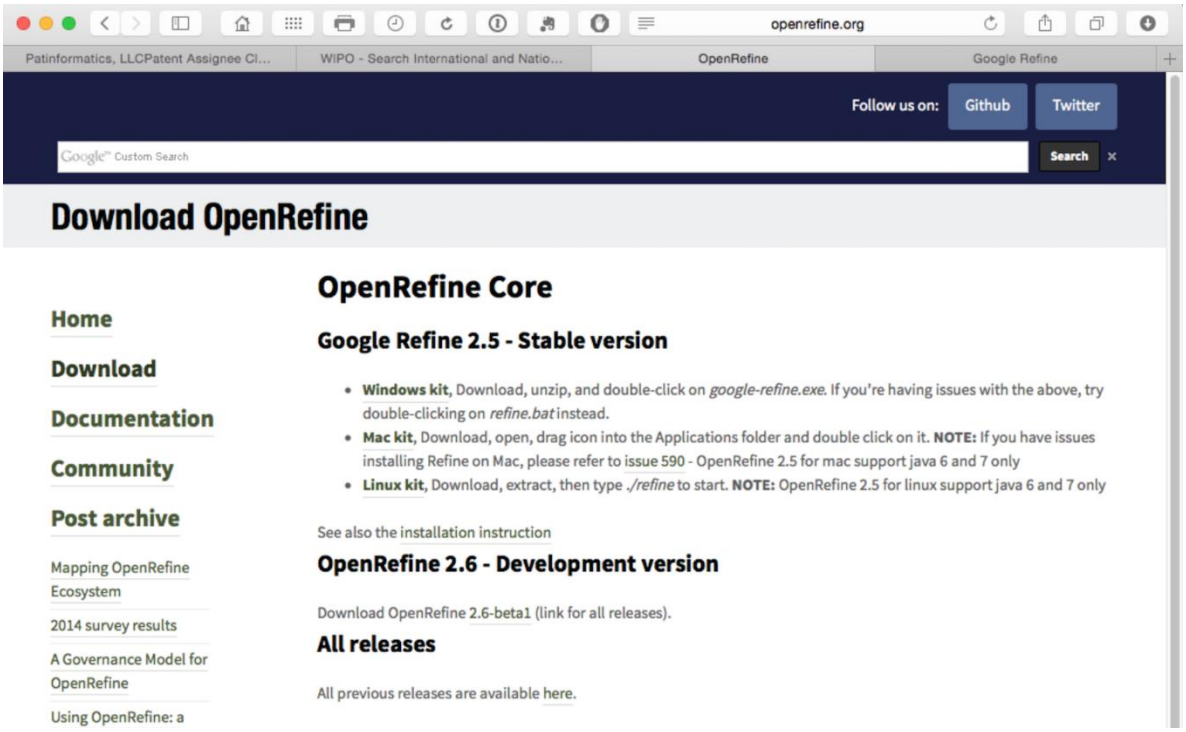
La razón por la que los analistas de patentes deberían utilizar Open Refine es que es la herramienta gratuita más fácil de usar y más eficiente para limpiar datos de patentes sin conocimientos de programación. Es muy superior a intentar las mismas tareas de limpieza en Excel u Open Office. La terminología que se usa puede tardar un tiempo en acostumbrarse, pero es posible desarrollar flujos de trabajo eficientes para limpiar y remodelar datos utilizando Open Refine y crear y reutilizar códigos personalizados necesarios para tareas específicas.

En este capítulo, utilizamos Open Refine para limpiar un conjunto de datos sin procesar de WIPO Patentscope que contiene casi 10,000 registros en bruto que hacen algún tipo de referencia a la palabra pizzaen todo el texto. Para seguir este capítulo utilizando uno de nuestros conjuntos de capacitación, descárguelo desde el repositorio de Github aquí o use su propio conjunto de datos.

8.1 Instalar Open Refine

Para instalar Open Refine, visite el sitio web de Open Refine y descargue el software para su sistema operativo:

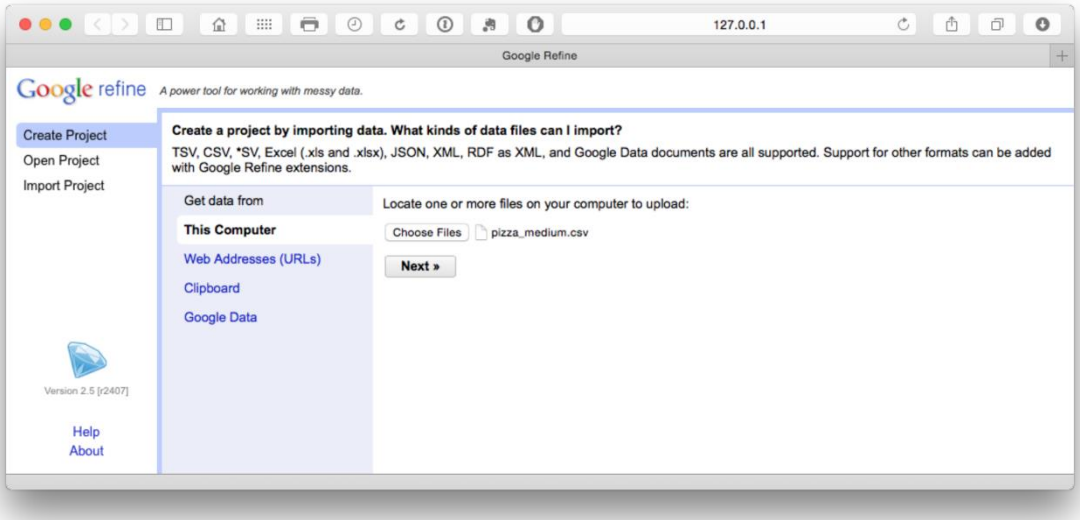
Desde la página de descarga seleccione su sistema operativo. Tenga en cuenta las extensiones hacia la parte inferior de la página para futuras referencias.



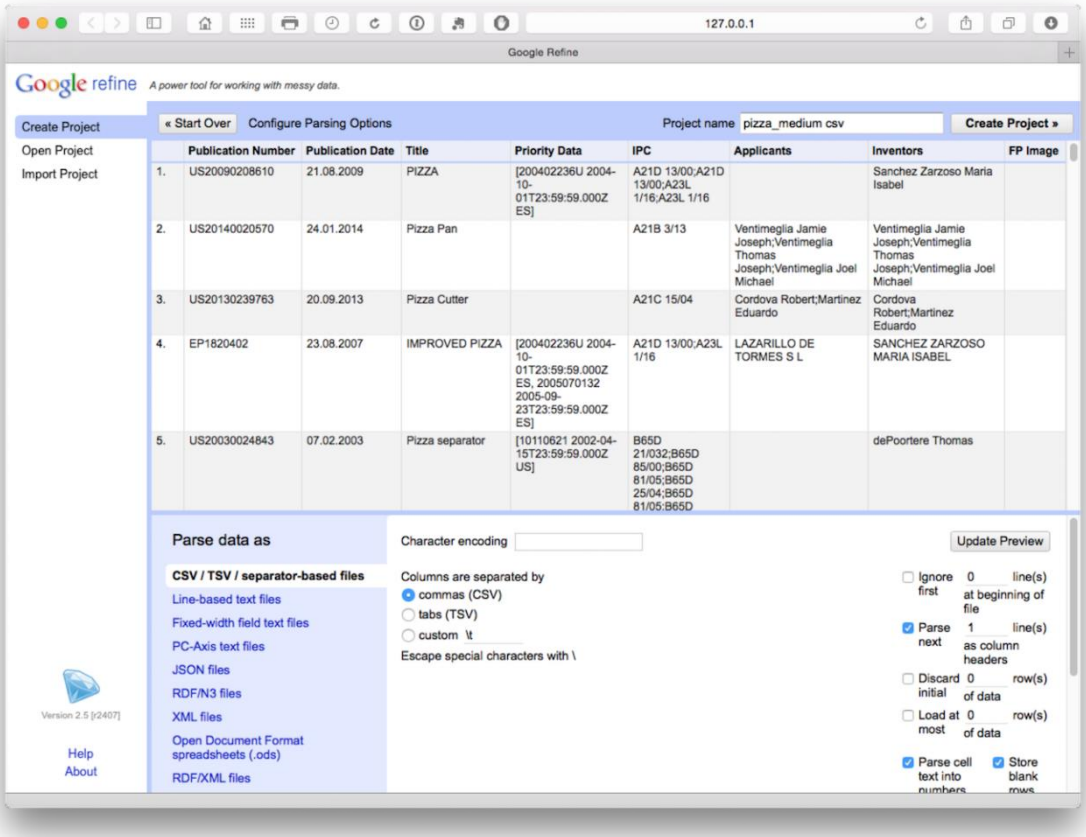
Al momento de escribir, cuando descarga Open Refine, en realidad se descarga e instala como Google Refine (reflejando su historial) y esa es la aplicación que deberá buscar y abrir.

8.2 Crea un proyecto

Utilizaremos el archivo Patentscope Pizza Medium que se puede descargar desde el repositorio aquí .



El archivo se cargará y luego intentará adivinar el separador de columna. Elija .csv.Tenga en cuenta que se puede importar una amplia gama de archivos y que hay opciones adicionales, como almacenar celdas en blanco como nulos que se seleccionan de forma predeterminada. En el conjunto de datos que cargará, hemos completado celdas en blanco con NAvalores para evitar posibles problemas al usar el relleno hacia abajo en Refinar abierto que se describe a continuación.



Haga clic **create Project** en la parte superior derecha de la barra como el siguiente paso.

8.3 Conceptos básicos de refinamiento abierto

Algunas características básicas de Open Refine pronto lo harán trabajar sin problemas. Aquí hay un recorrido rápido.

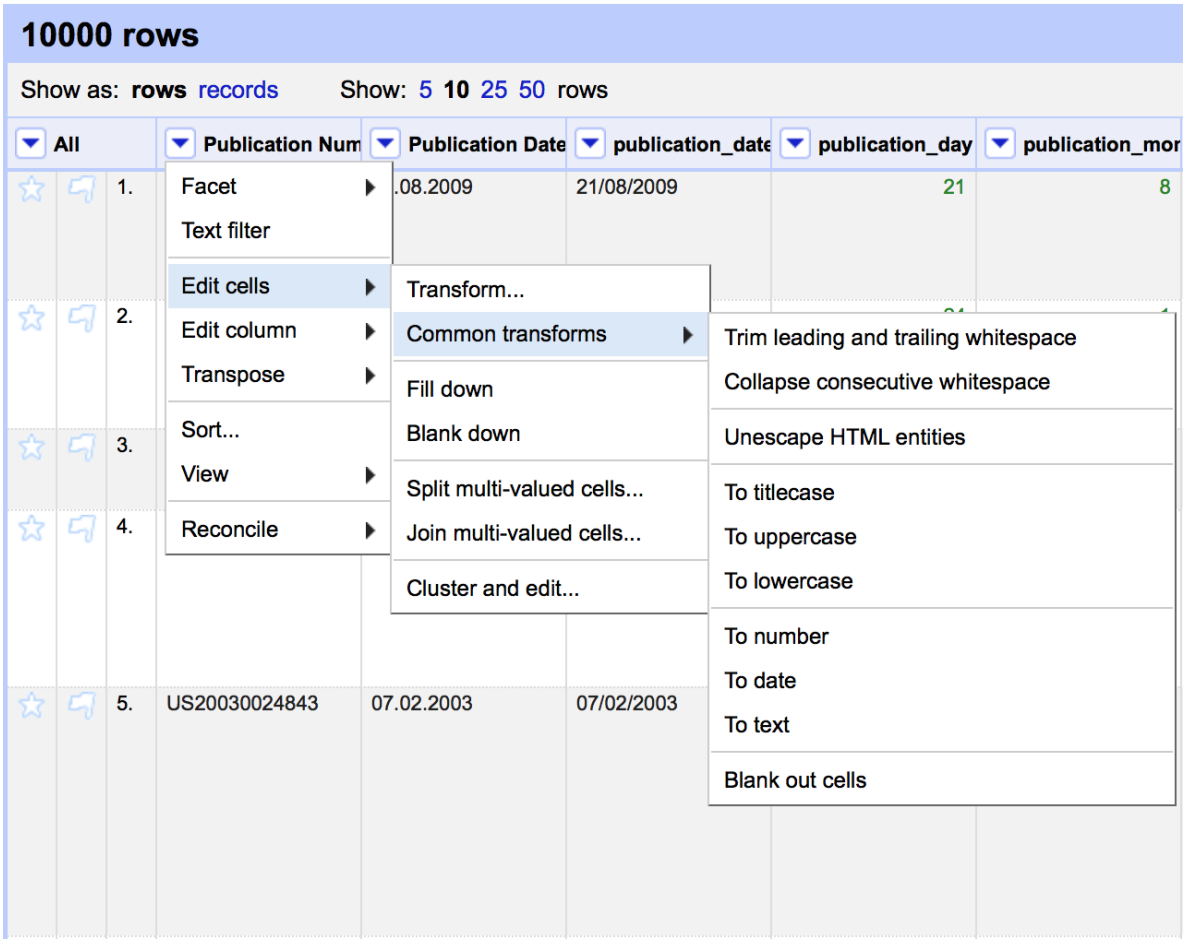
8.3.1 Abrir Refinar se ejecuta en un navegador

Open Refine es una aplicación que vive en su computadora, pero se ejecuta en un navegador. Sin embargo, no requiere una conexión a Internet y no pierde su trabajo si cierra el navegador.

8.3.2 Abrir Refinar trabajos en columnas.

En la parte superior de cada columna hay un menú desplegable. Esté listo para usar estos menús bastante. En particular, a menudo utilizará el Edit cells > Common

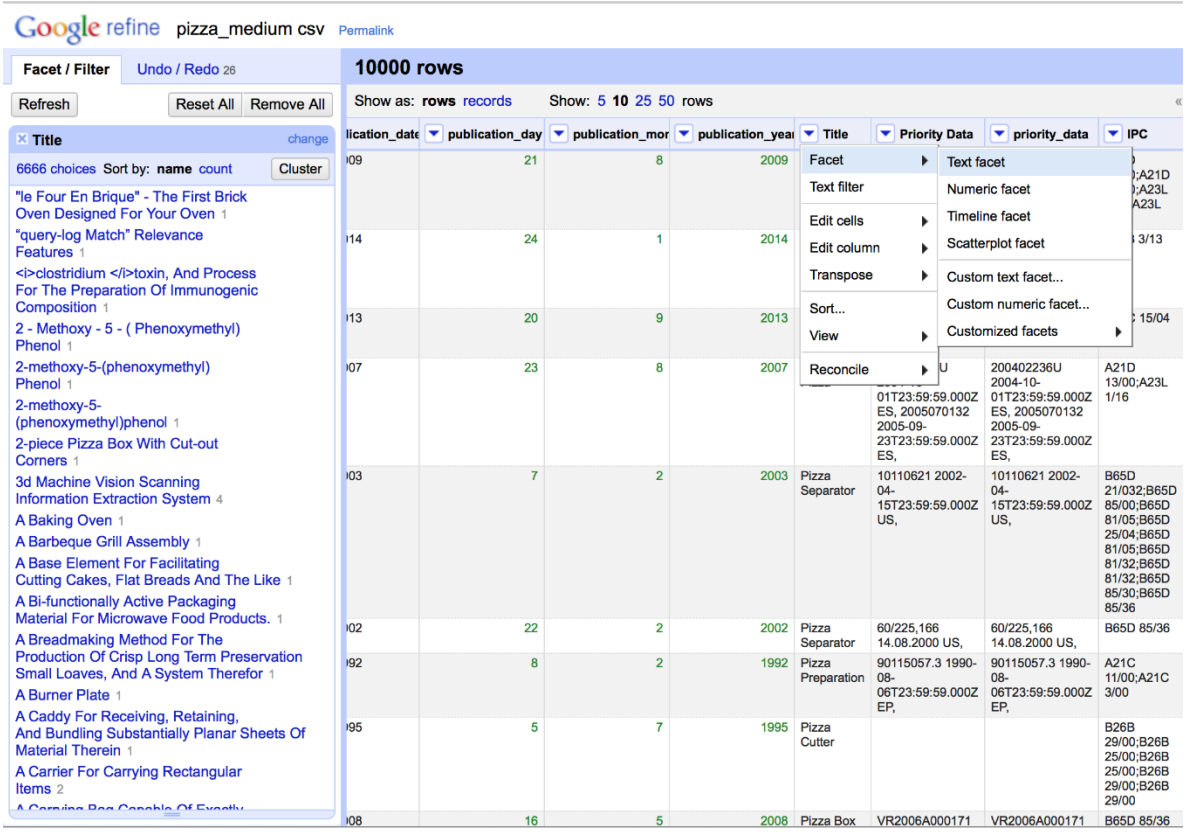
tranforms que se muestra a continuación para funciones como el recorte de espacios en blanco.



Otros menús importantes son Edit column, inmediatamente a continuación Edit cells, para copiar o dividir columnas en nuevas columnas.

8.3.3 Abrir Refinar trabajos con facetas.

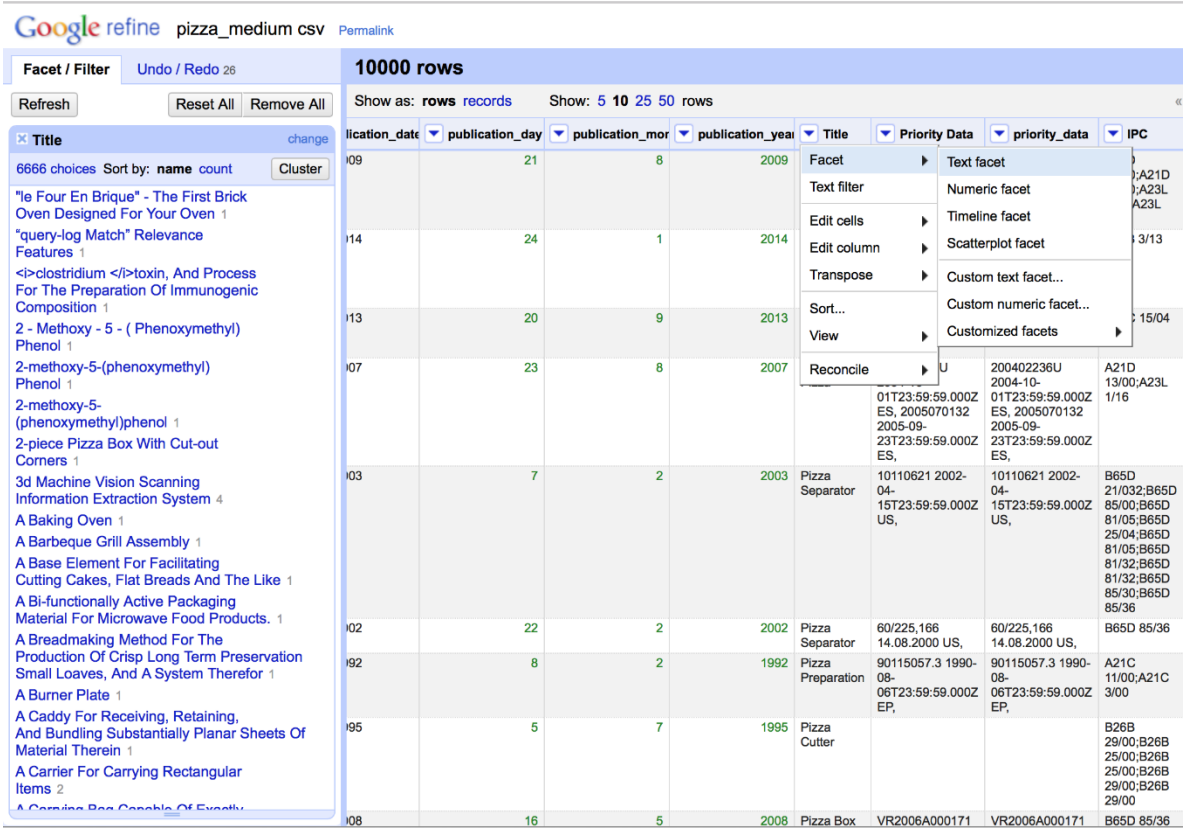
El término facet puede ser confuso inicialmente, pero básicamente abre una ventana que organiza los elementos en una columna para su inspección, clasificación y edición, como podemos ver a continuación. Esto es importante porque es posible identificar problemas y abordarlos. También es posible aplicar una variedad de algoritmos de agrupación en clústeres para limpiar los datos. Tenga en cuenta que el tamaño de la ventana de faceta se puede ajustar arrastrando la parte inferior de la ventana como lo hemos hecho en esta imagen.



Al pasar el cursor sobre un elemento dentro de la ventana de faceta, aparece un pequeño botón edit que permite editar, como eliminar *y* del título.

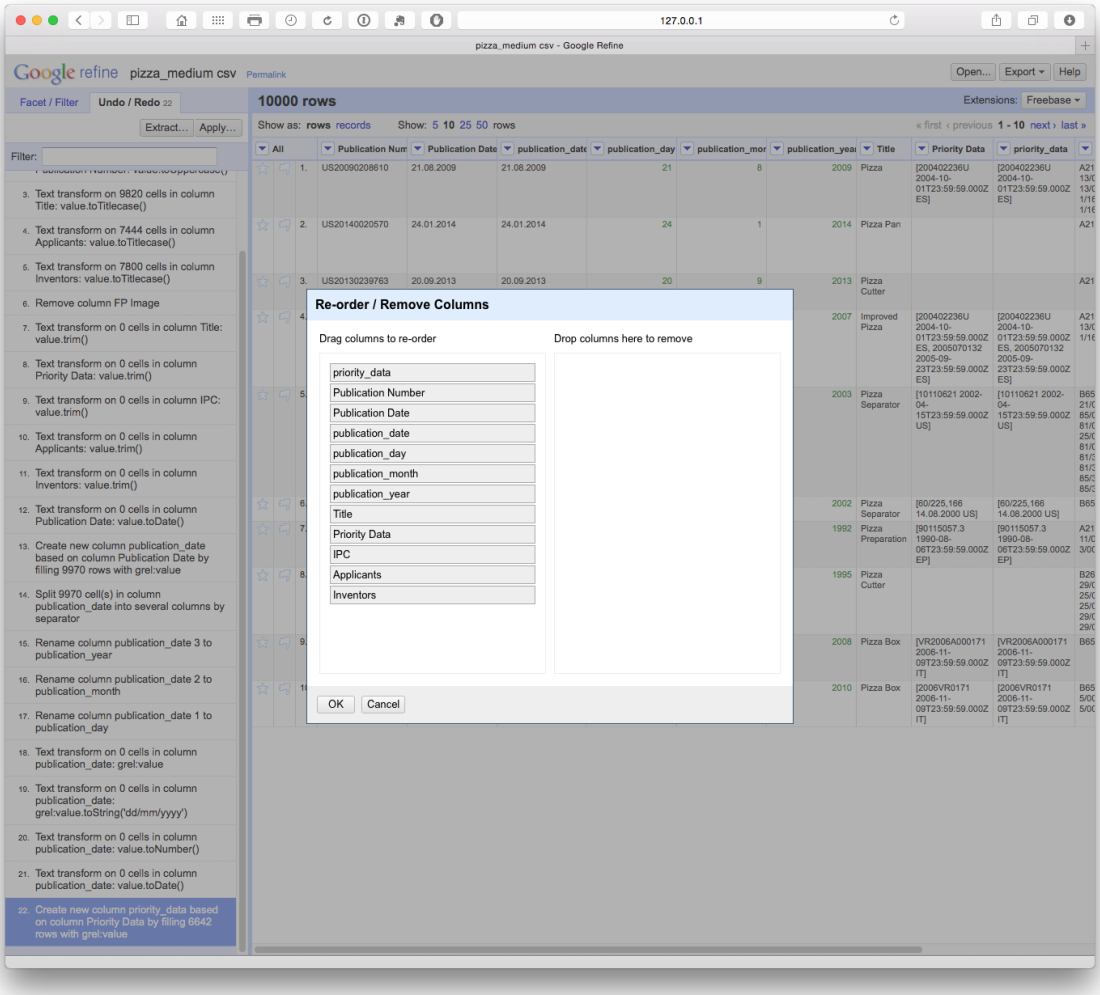
8.3.4 Facetas personalizadas

El siguiente menú de facetas muestra un menú personalizado con una gama de opciones. Al seleccionar Custom text facet (ver a continuación) aparece una ventana emergente que permite el uso del código en Open Refine Expression Language (GREL) para realizar tareas que no están cubiertas por los elementos del menú principal. Este lenguaje es bastante simple y puede abarcar desde fragmentos breves para buscar y reemplazar texto hasta funciones más complejas que se pueden reutilizar en el futuro. Demostraremos el uso de esta función a continuación.



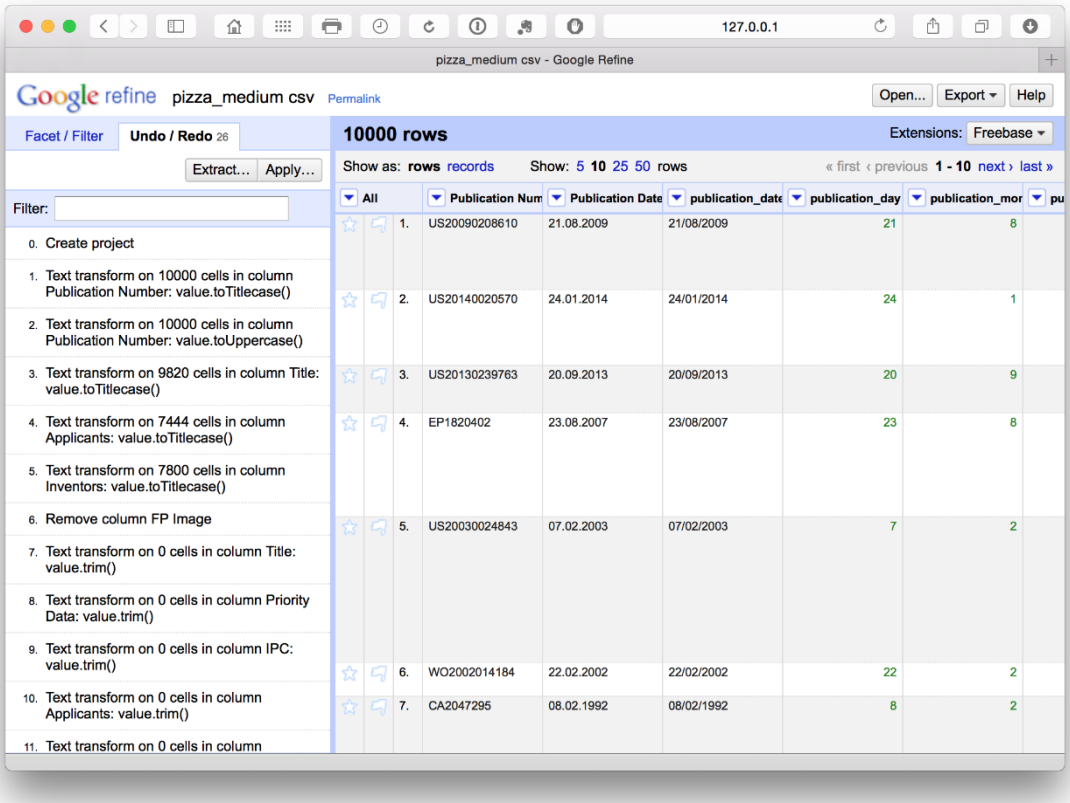
8.3.5 Reordenar columnas

Hay dos opciones para reordenar columnas. Lo primero es seleccionar el menú de la columna luego Edit column > Move column to beginning. La segunda opción, que se muestra a continuación, es seleccionar el Allmenú desplegable en la primera columna y luego Edit columns > Re-order/remove columns. En el menú emergente de campos, arrastre el campo deseado a la parte superior de la lista. En este caso, hemos arrastrado la priority_datecolumna a la parte superior de la lista. Ahora aparecerá como la primera columna de datos.



8.3.6 Deshacer y rehacer

Open Refine realiza un seguimiento de cada acción y le permite retroceder varios pasos o al principio. Esto es particularmente útil cuando se prueba si un enfoque particular para la limpieza (por ejemplo, dividir columnas o usar un fragmento de código) satisfará sus necesidades. En particular, significa que puede explorar y probar enfoques sin preocuparse por perder su trabajo anterior.



Sin embargo, puede ser importante planificar los pasos en su operación de limpieza para evitar problemas en etapas posteriores. Puede ser útil usar un bloc de notas como lista de verificación (ver más abajo). El principal problema que puede surgir es cuando la limpieza avanza varios pasos sin completarse completamente en un paso anterior. En algunos casos, esto puede requerir volver a ese paso anterior, reiniciar y repetir pasos anteriores. A medida que se familiarice con Open Refine, será más fácil elaborar una secuencia adecuada para su flujo de trabajo.

8.3.7 Exportando

Cuando se completa un ejercicio de limpieza, se puede exportar un archivo en una variedad de formatos. Cuando trabaje con datos de patentes, espere crear más de un archivo (por ejemplo, núcleo, solicitantes, inventores, IPC) para permitir el análisis de aspectos de los datos en otras herramientas. En este capítulo crearemos dos archivos.

1. Una versión limpia de los datos originales.
2. Un archivo de solicitantes que separa los datos por cada solicitante.

8.4 Limpieza básica

Este es el primer paso para trabajar con un conjunto de datos y tendrá sentido realizar algunas tareas básicas de limpieza antes de continuar. El conjunto de datos de Pizza Medium con el que estamos trabajando en este artículo es en bruto en el sentido de que la única limpieza hasta ahora ha sido eliminar las dos filas vacías en la cabecera de la tabla de datos y rellenar las celdas en blanco con valores de NA. La razón por la que tiene sentido realizar una limpieza básica antes de trabajar con los datos del solicitante, el inventor o el IPC es que este proceso generará nuevos conjuntos de datos.

Tenga en cuenta que Open Refine no es el programa más rápido y asegúrese de asignar tiempo suficiente para las tareas de limpieza y de que esté preparado para ser paciente mientras el programa ejecuta algoritmos para procesar los datos. Tenga en cuenta que Open Refine guardará su trabajo y puede volver a él más tarde.

Cuando se trabaja con Open Refine, normalmente trabajaremos en una columna a la vez. Sin embargo, la clave checklist para los pasos de limpieza es:

1. Asegúrese de tener una copia de seguridad del archivo original. Crear un archivo .zip y marcarlo con el nombre raw puede ayudar a conservar el original.
2. Abra y guarde un archivo de texto code book para anotar los pasos tomados para limpiar los datos (por ejemplo, pizza_codebook.txt).
3. Regularice los caracteres (por ejemplo, título, minúsculas, mayúsculas).
4. Quite los espacios en blanco iniciales y finales.
5. Codificación de direcciones y problemas relacionados.

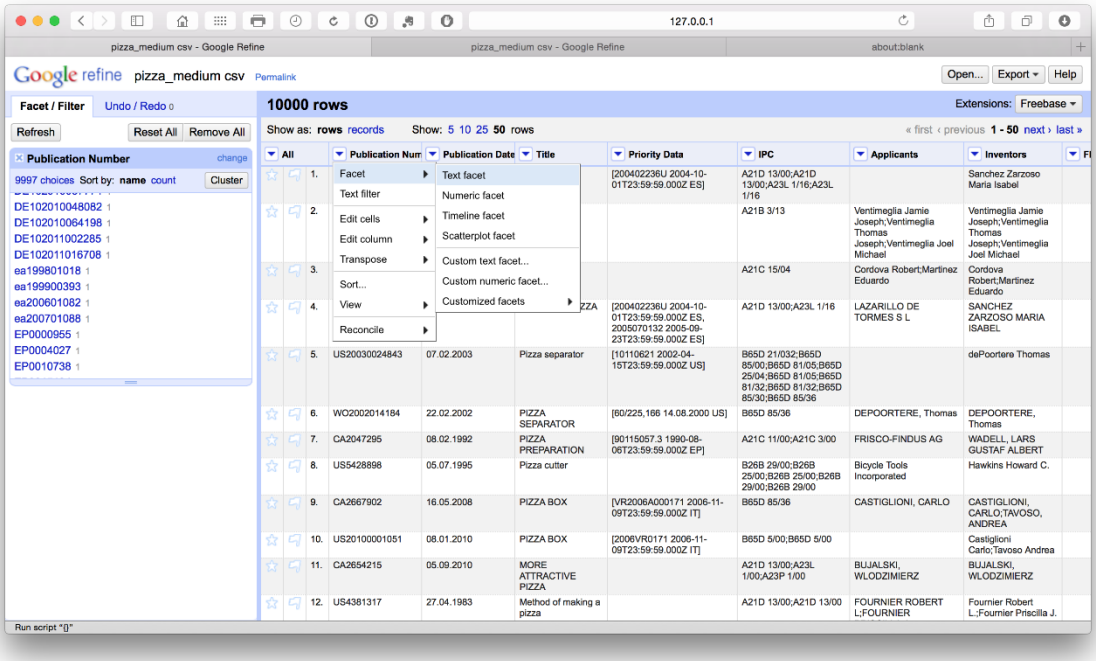
Acciones adicionales:

6. Transformar fechas
7. Acceda a información adicional y cree nuevas columnas y / o filas.

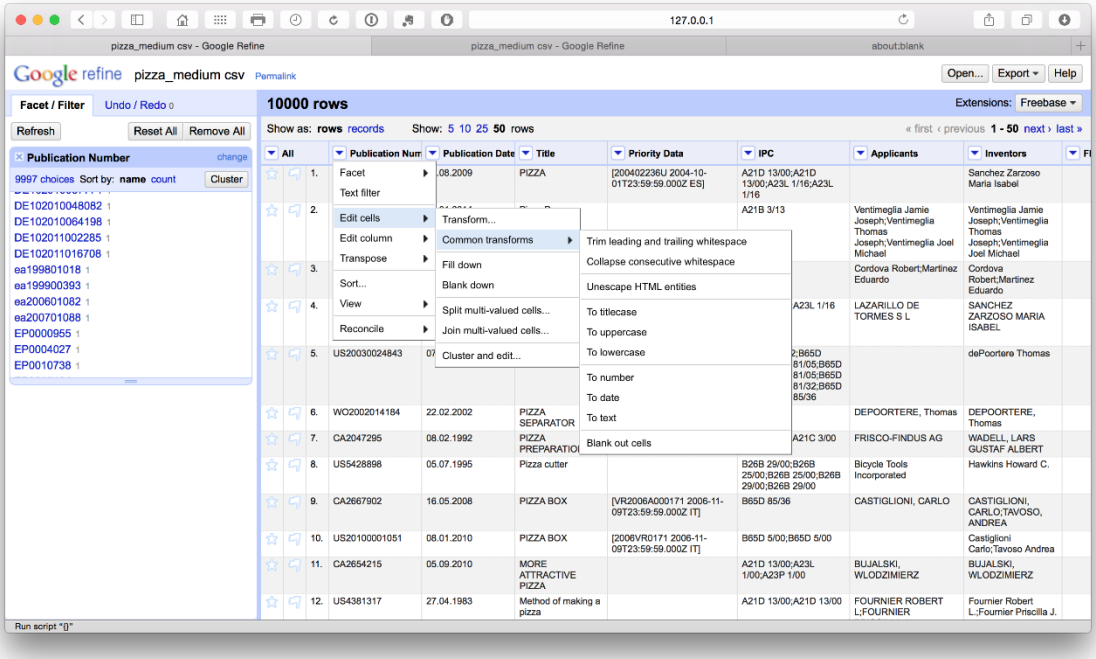
En general, abordaremos estas tareas en cada columna y los pasos 6 y 7 no siempre se aplicarán. La creación de un libro de códigos le permitirá mantener una nota de todos los pasos tomados para limpiar un conjunto de datos. El libro de códigos debe guardarse con el conjunto de datos limpiado (por ejemplo, en la misma carpeta) como punto de referencia si necesita realizar más trabajos o si los colegas desean comprender los pasos de la transformación.

8.4.1 Cambio de caja

La primera columna de nuestro conjunto de datos de pizza de Patentscope es el número de publicación. Para inspeccionar lo que está sucediendo y debe limpiarse en esta columna, primero seleccionaremos el menú de la columna y elegiremos text facetdel menú desplegable. Esto generará el panel de menú lateral que podemos ver a continuación que contiene los datos. Entonces podemos inspeccionar la columna por problemas.



Cuando nos desplazamos hacia abajo en el panel lateral, podemos ver que algunos números de publicación tienen un código de país en minúsculas (en este caso, en ea lugar de EA para la Organización de Patentes de Eurasia que utiliza los códigos de país estandarizados dela OMPI). Para abordar esto seleccionamos el menú de la columna Edit cells > Common transforms > to Uppercase.



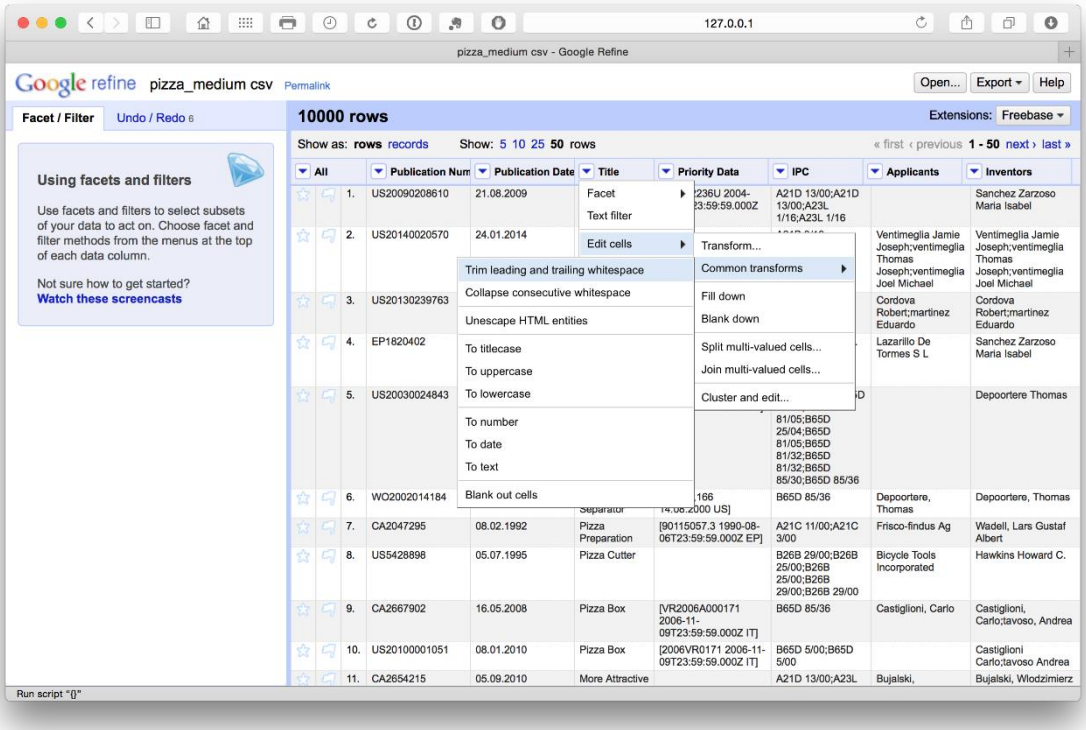
Si nos desplazamos hacia abajo, todos los números de publicación se habrán convertido en mayúsculas. Esto facilitará la extracción de los códigos de país de publicación en una etapa posterior.

8.4.2 Regularizar caso

Para las otras columnas de texto, es conveniente repetir el paso de las transformaciones comunes y seleccionar to titlecase. Tenga en cuenta que esto generalmente funcionará bien para el campo de título, pero puede que no siempre funcione tan bien en campos concatenados como los nombres de los solicitantes y los inventores. Repita este paso después de la separación de estos campos concatenados (consulte a continuación sobre los solicitantes). Si el resumen o las reclamaciones estuvieran presentes, no regularizaríamos esos campos de texto.

8.4.3 Eliminar los espacios en blanco iniciales y finales

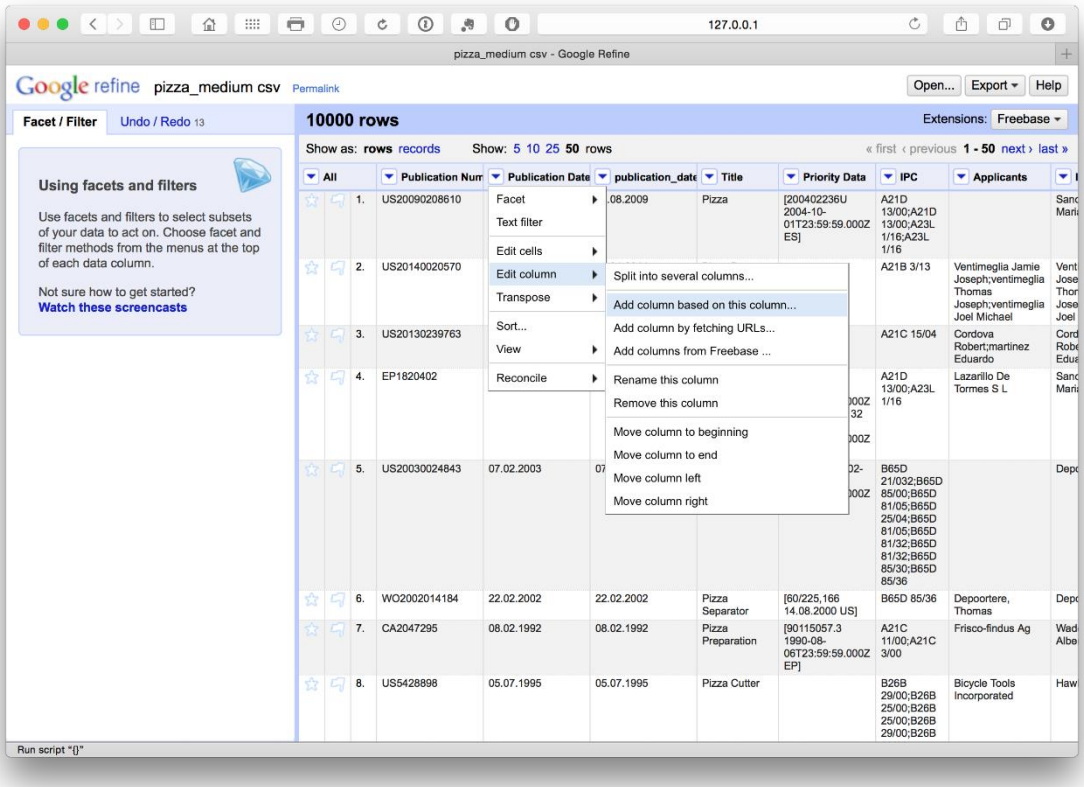
Para eliminar los espacios en blanco iniciales y finales en una columna, seleccionamos Edit cells > Common transforms > Trimlos espacios en blanco iniciales y finales en las columnas.



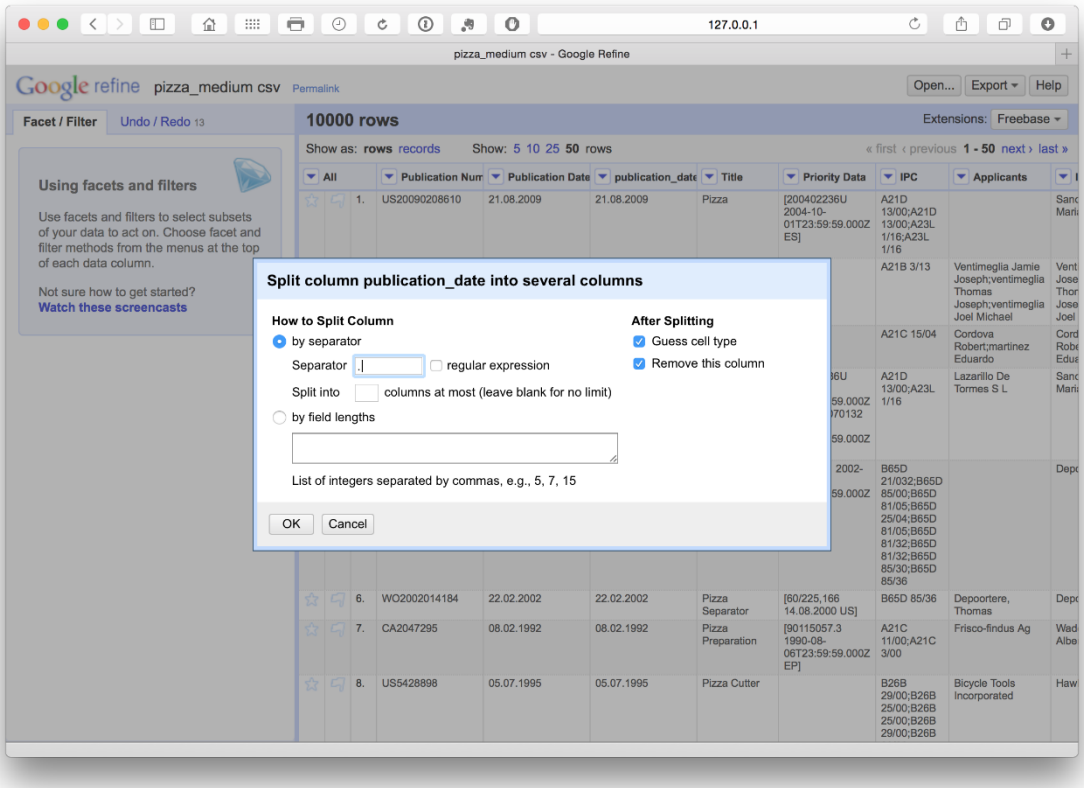
Tenga en cuenta que después de dividir las celdas concatenadas con múltiples entradas, como los campos de los solicitantes y los inventores, es una buena idea repetir el ejercicio de recorte cuando se complete el proceso para evitar posibles espacios en blanco al inicio de las nuevas entradas de nombres.

8.4.4 Añadir columnas

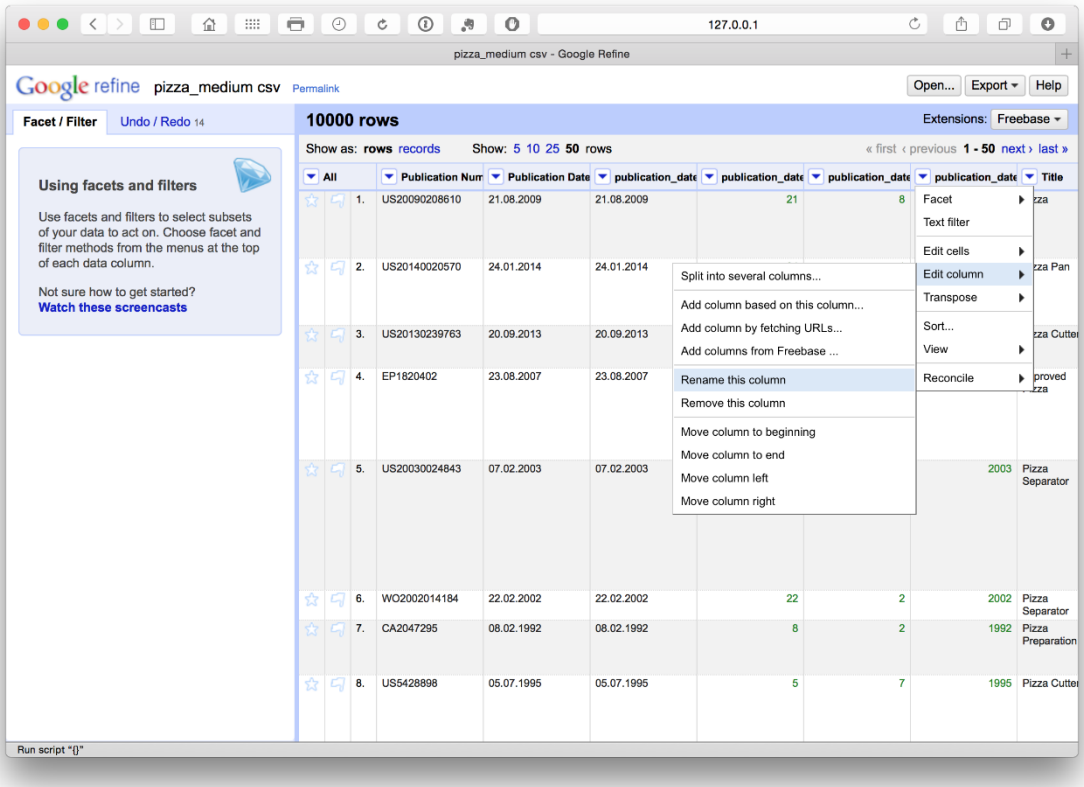
También podemos agregar columnas seleccionando el menú de columnas y Edit Column > Add column based on this column. En este caso, hemos agregado una columna llamada fecha_de_publicación.



Tenemos una serie de opciones con respecto a las fechas (ver más abajo). En este caso, queremos separar la información de la fecha en columnas separadas. Para ello podemos utilizar Edit column > Split into several columns. También podemos elegir el separador para la división, en este caso. y si mantener o eliminar la columna de origen. En este caso, seleccionamos conservar la columna original y creamos tres nuevas columnas relacionadas con la fecha. Podríamos, según sea necesario, eliminar la columna de fecha y mes si solo necesitáramos el campo del año.



Luego podemos cambiar el nombre de estas columnas usando la edición Edit Column > Rename this column. Tenga en cuenta que, en este caso, el uso de minúsculas y guiones bajos marca las columnas que estamos creando o editando como un indicador para uso interno que nos informa que esta es una columna que hemos creado. En una etapa posterior, cambiaremos el nombre de los campos originales para marcarlos como originales.



8.4.5 Codificación de direcciones y problemas relacionados

La sección de Recetas de la documentación proporciona consejos útiles y código de ejemplo para tratar con la codificación y los problemas relacionados que se reproducen aquí:

1. personajes corruptos. Esto surge de la agregación de datos de diferentes fuentes. En Patentscope los datos se convierten a UTF8. Sin embargo, si se encuentran problemas, seleccione Edit cells > Transforme intente ingresar lo siguiente.

```
value.reinterpret("utf-8")
```

Puede ser necesario explorar y probar otros conjuntos que se pueden identificar aquí .

2. Escape de caracteres html / XML, por ejemplo, & amp

La fuente más probable de datos de patentes es XML, pero para estar seguro, lo siguiente debería escapar (eliminar) el código html y XML que aparece en el texto.

```
value.unescape("html").unescape("xml")
```

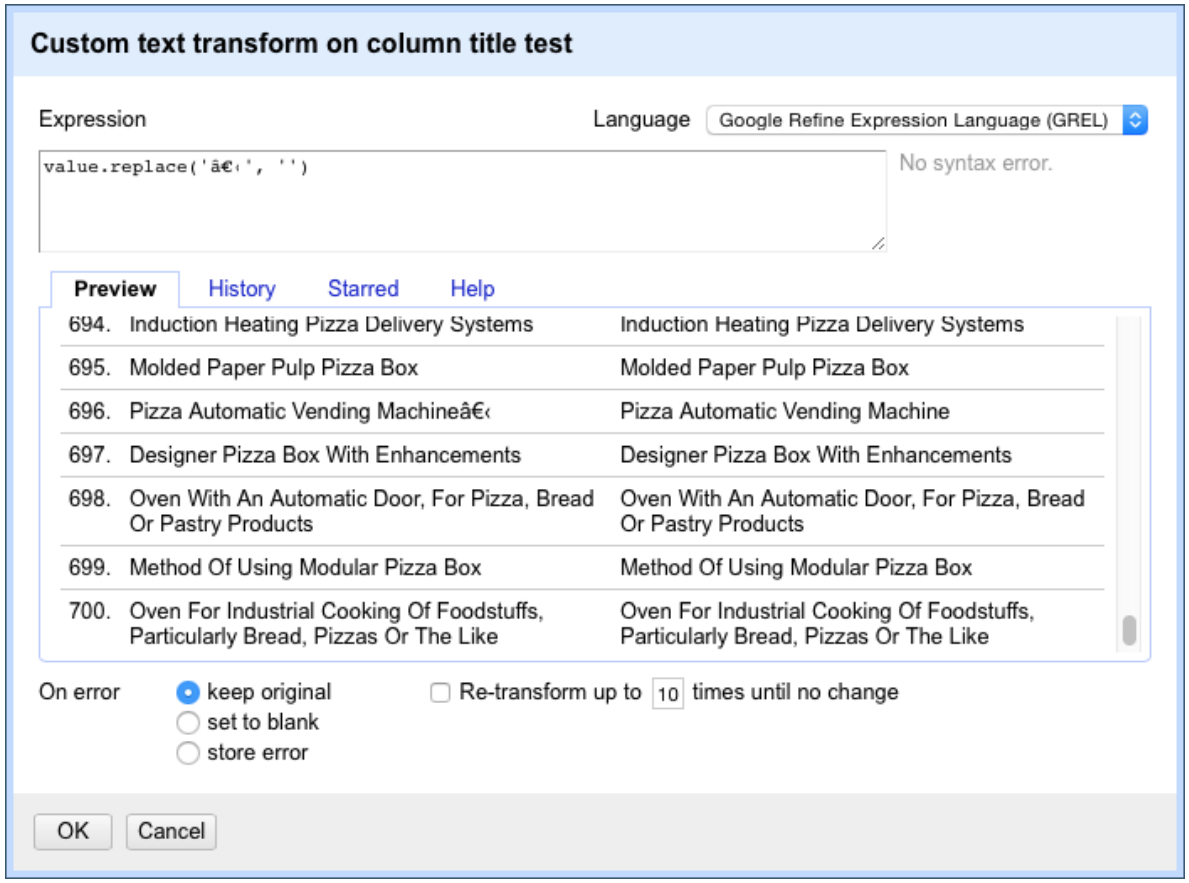
3. Signos de interrogación

Los signos de interrogación a menudo aparecen para que los caracteres que no se pueden representar son un signo de problemas de codificación. Además, los espacios sin interrupción se pueden representar como (Unicode (16)). Para encontrar un valor Unicode, vaya a Edit cells > Transforme ingrese Unicode (valor) que transformará todos los caracteres en números Unicode. Desde allí puedes buscar el problema.

Se propone una solución rápida en la documentación que puede funcionar en algunas circunstancias.

```
split(escape(value,'xml')," ")[0]
```

Dentro de este conjunto de datos en particular, encontramos que estos consejos rápidos no funcionaron (probablemente porque el texto ya se había convertido a UTF-8). Sin embargo, si todo lo demás falla, una alternativa es simplemente buscar y reemplazar en Transformar como en el ejemplo a continuación.

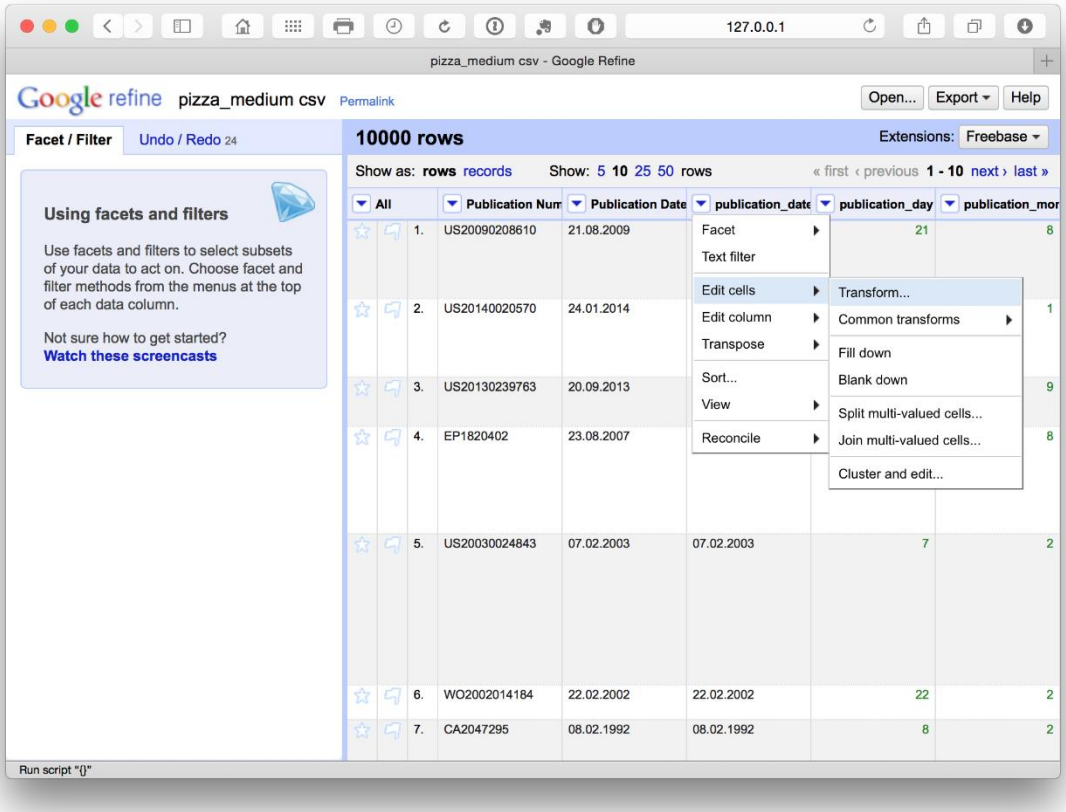


Esta no es una solución muy satisfactoria porque requiere la inspección del conjunto de datos para identificar los problemas de caracteres específicos y luego reemplazar el valor. Eso llevará mucho tiempo.

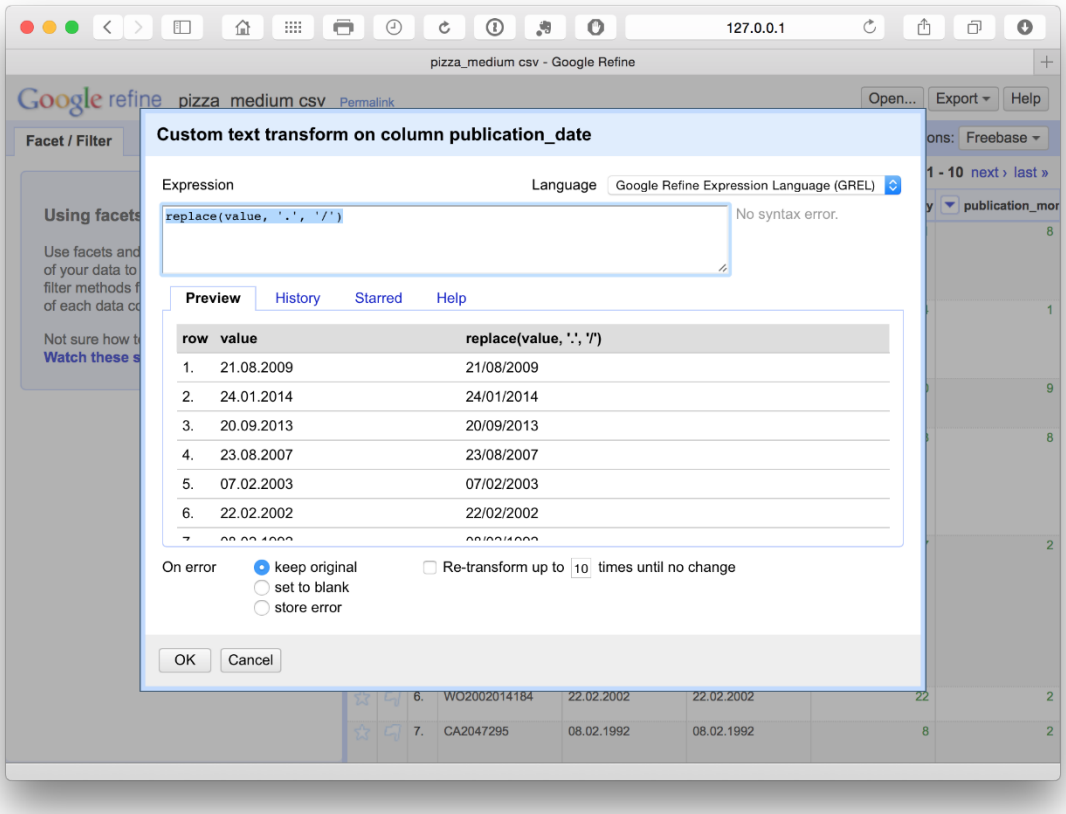
8.4.6 Reformateo de fechas

Un problema que podemos encontrar es que la definición de fecha estándar en los documentos de patente (por ejemplo, 21.08.2009) puede no ser reconocida como un campo de fecha en nuestro software de análisis porque las fechas pueden ser ambiguas desde la perspectiva del código del software. Por ejemplo, ¿cómo se debe interpretar el 12/08/2009 o el 12/08/2009?

Alternativamente, como en este caso, los puntos decimales pueden no interpretarse correctamente como una fecha en algún software (por ejemplo, R). Podríamos anticipar esto y transformar los datos en una forma más reconocible, como el 21/08/2009. Una forma muy sencilla de hacerlo es mediante el uso de una función de reemplazo. En este caso seleccionamos el menú para el publication_datecampo y luego Edit cells > Transform.



Esto produce un menú donde ingresamos un código de reemplazo GREL simple.




```
replace(value, '.', '/')
```

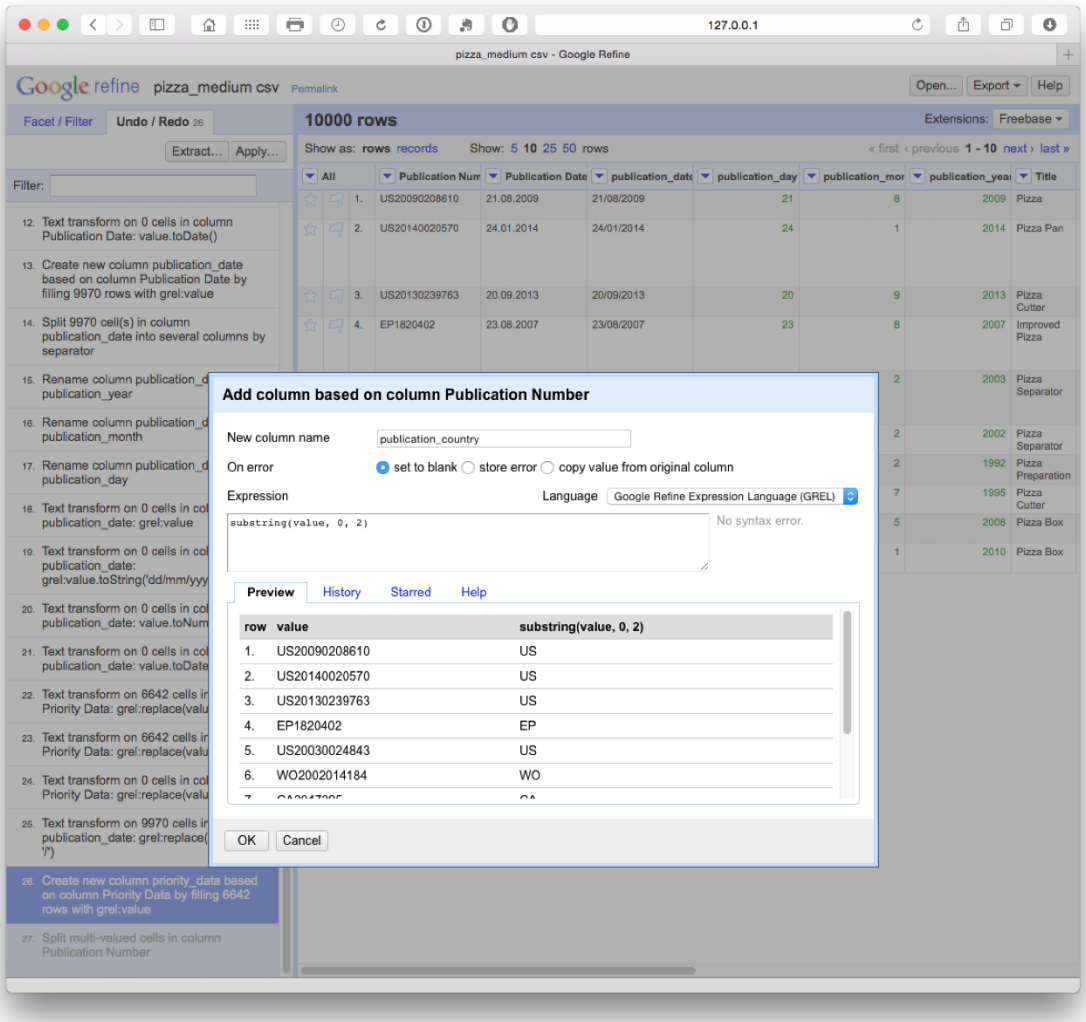
Este código de reemplazo simple es básicamente el mismo que buscar y reemplazar en Excel u Open Office. Además, podemos ver las consecuencias de la elección en el panel antes de ejecutar el comando. Esto es extremadamente útil para detectar problemas. Por ejemplo, si intentamos dividir un campo en una coma, podemos descubrir que hay varias comas en una celda (vea el Priority Datacampo para esto). Al probar el código en el panel, podríamos trabajar para encontrar una solución o editar los textos ofensivos en el panel de facetas principal.

Para encontrar otros códigos simples, visite la página Abrir recetas de refinamiento .

Ahora nos centraremos en extraer información de algunas de las columnas antes de guardar el conjunto de datos y seguir adelante.

8.4.7 Acceso a información adicional

Hay una variedad de piezas de información que están ocultas en los datos dentro de las columnas. Por ejemplo, los datos de Patentscope no contienen un campo de país de publicación. En particular, tenga en cuenta que Patentscope combina todas las publicaciones de un registro de aplicación en un expediente. Por lo tanto, solo estamos viendo un registro para un conjunto de documentos (en Patentscope se puede acceder al expediente más amplio de un registro a través de la sección Documentos del sitio web). Esto es muy útil para reducir la duplicación, pero es importante tener en cuenta que no estamos viendo a la familia más amplia en nuestra tabla de datos. Sin embargo, podemos trabajar con la información al principio del número de publicación en los registros de Patentscope utilizando un código muy simple y crear una nueva columna (como la anterior) basada en los valores que se muestran a continuación.

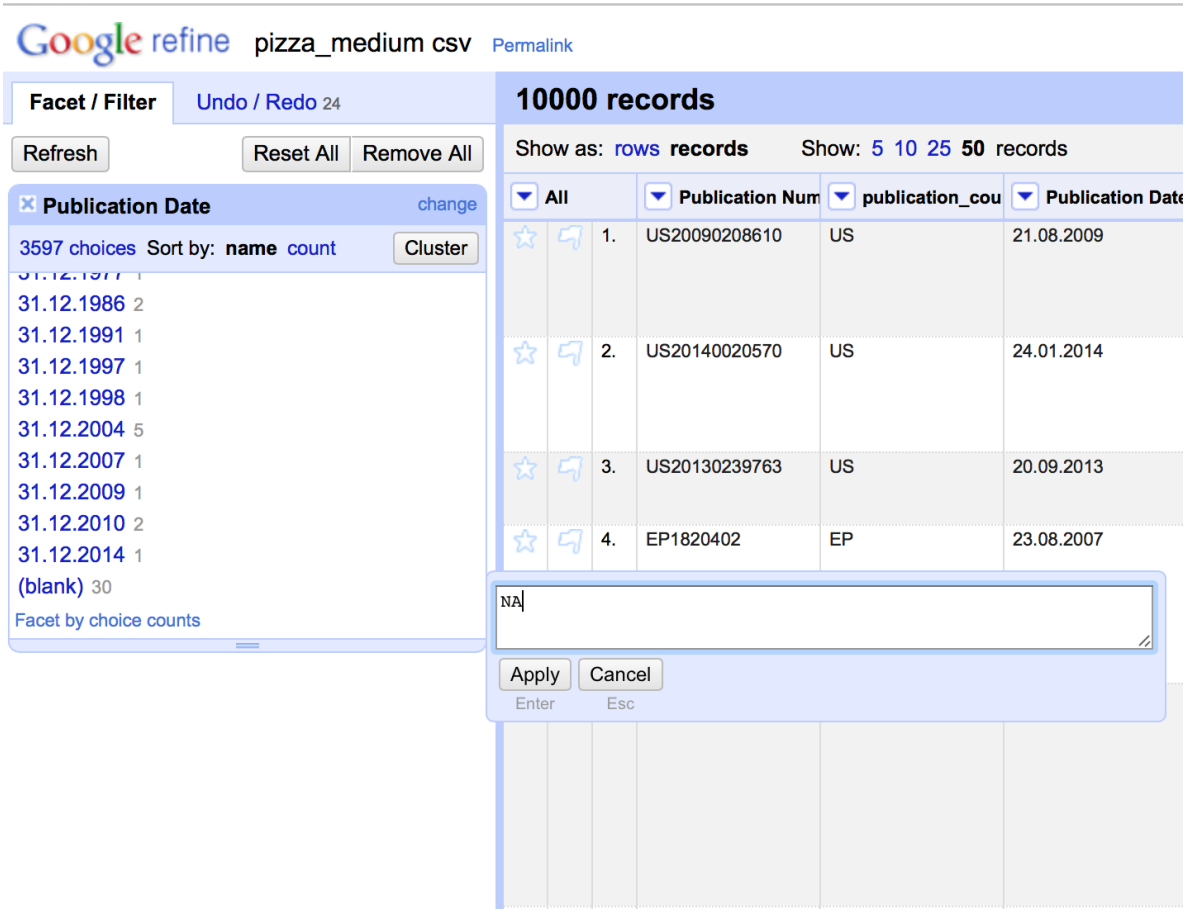



```
substring(value, 0, 2)
```

Tenga en cuenta que el código comienza a contar desde 0 (por ejemplo, 0, 1 = U, 2 = S). La primera parte del código se ve en el campo de valor. 0 le dice al código que comience a contar desde 0 y el 2 le dice que lea los dos caracteres de 0. Podríamos cambiar estos valores, por ejemplo, a 1 y 4 para capturar solo una parte de un número.

8.5 Rellenar celdas en blanco

El llenado de celdas en blanco con un valor (NA para No disponible) para evitar problemas de cálculo con las herramientas de análisis más adelante puede realizarse seleccionando cada columna, creando una faceta de texto, desplazándose hacia abajo hasta la parte inferior de la faceta seleccionando (blank), editando y luego ingresando NA para el valor.



Tenga en cuenta que esto requiere algo de tiempo (hasta que se encuentra un método más rápido) pero tiene la ventaja de ser preciso. Generalmente, es más rápido abrir el archivo en Excel (u Open Office) y usar buscar y reemplazar con el cuadro de búsqueda dejado en blanco y NA en el campo de reemplazo en la tabla de datos. Por esa razón, las celdas en blanco no deben aparecer en el conjunto de

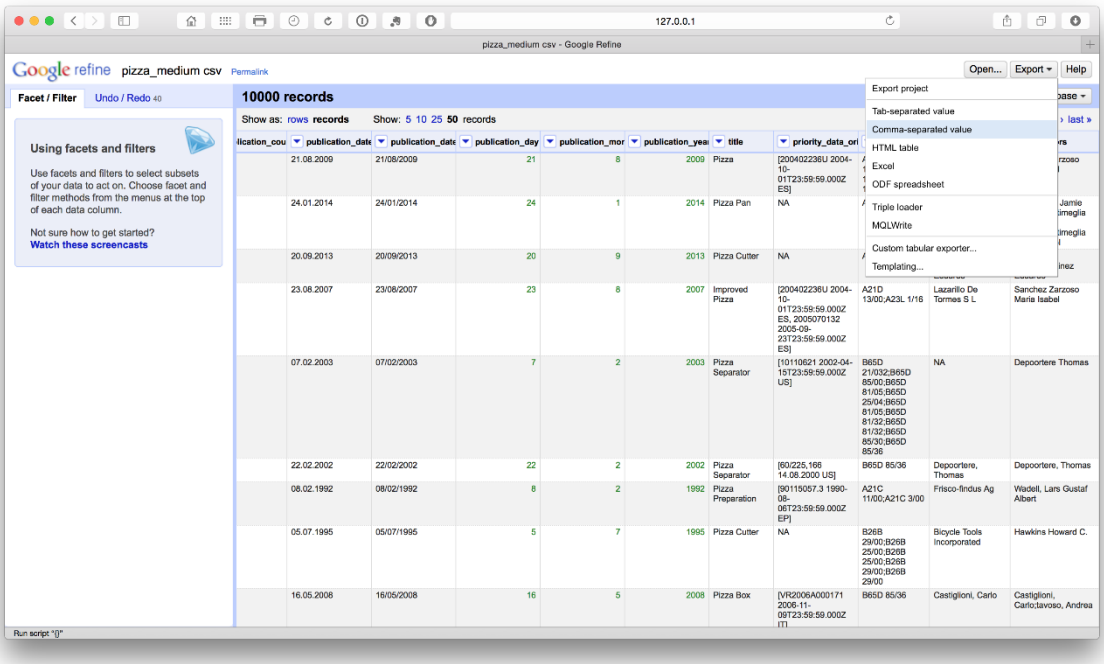
datos que está utilizando en este artículo. Sin embargo, en los pasos posteriores a continuación, generaremos celdas en blanco al dividir el campo del solicitante y el inventor. Por eso es importante conocer este procedimiento.

8.6 Renombrando columnas

En esta etapa, tenemos un conjunto de columnas que se mezclan entre el caso de la oración original y las adiciones en minúsculas sin espacios como publication_number. Esta es una cuestión de preferencia personal, pero en general es una buena idea regularizar el caso de todas las columnas para que sean fáciles de recordar. En este caso, también agregaremos la palabra originala las columnas para distinguir entre aquellos creados al limpiar los datos y aquellos que hemos creado.

8.7 Exportación de datos

Cuando estamos contentos de haber trabajado con los pasos de limpieza del núcleo, es una buena idea exportar el nuevo conjunto de datos del núcleo. Es importante hacer esto antes de los pasos que se describen a continuación, ya que conserva una copia del conjunto de datos central que se puede usar para la separación (o actividades de división) en solicitantes, inventores, IPC, etc., durante los próximos pasos. Es importante que este clean conjunto de datos esté tan limpio como sea razonablemente posible antes de continuar. La razón de esto es que ***cualquier ruido o problema se multiplicará*** cuando pasemos a los siguientes pasos. Esto puede requerir una repetición importante de los pasos de limpieza en los archivos posteriores creados para solicitantes o inventores. Por lo tanto, asegúrese de estar contento de que los datos estén tan limpios como sea razonablemente posible en esta etapa. Luego elija exportar desde el menú y el formato deseado (preferiblemente. csv o .tab si usa herramientas de análisis más adelante).



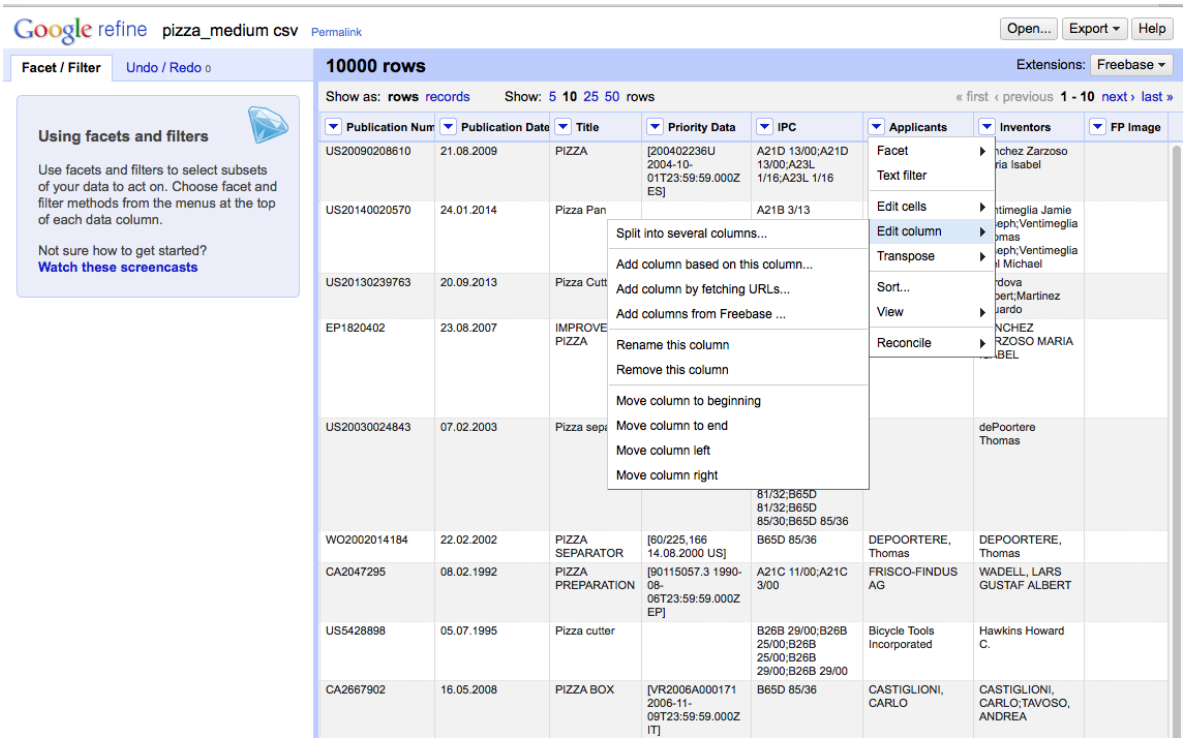
8.8 dividir los solicitantes

Este artículo se inspiró en un tutorial muy útil sobre cómo limpiar nombres de asignatarios con Google Refine por Anthony Trippe. Anthony también es el autor de las próximas Directrices de la OMPI para la preparación de informes de patentes de paisaje y el sitio web Patentinformatics LLC ha desempeñado un papel pionero en la promoción del análisis de patentes. Seguiremos este ejemplo utilizando nuestro conjunto de datos de patentes de pizzas para que pueda visualizarse en una variedad de herramientas. Si aún no lo ha hecho, puede descargar el conjunto de datos aquí .

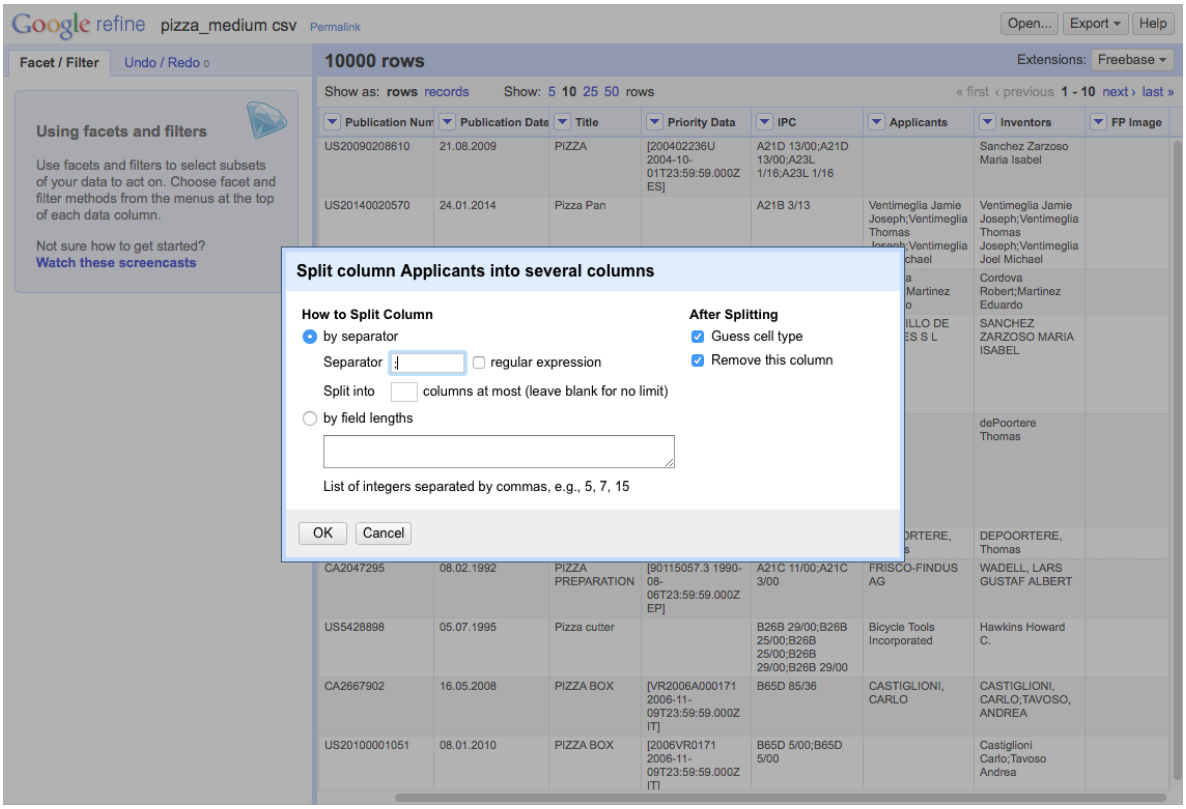
En realidad, tenemos dos opciones aquí y las revisaremos para que pueda resolver sus necesidades en una situación particular. Tenga en cuenta que puede usar Deshacer en Google Refine para volver al punto inmediatamente antes de probar estos enfoques. Sin embargo, si ha seguido los pasos de limpieza de datos anteriores, asegúrese de que ya ha exportado una copia del conjunto de datos principal.

8.8.1 Situación 1 - Primeros solicitantes

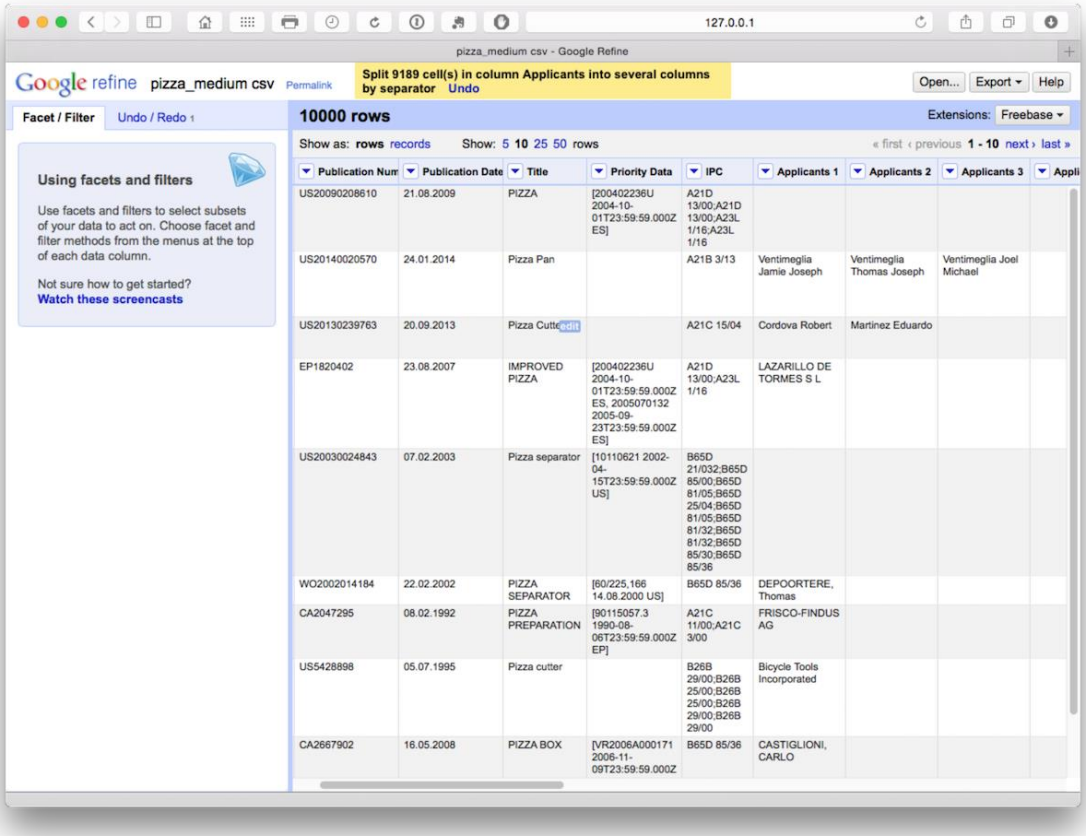
Como se discutió por Anthony Tripp podríamos dividir la columna de la solicitante en columnas separadas por la elección, Applicants > Edit column > Split into several columns.



Entonces tenemos que seleccionar el separador. En este caso (y normalmente con datos de patente), es ;.



Esto producirá un conjunto de 18 columnas.

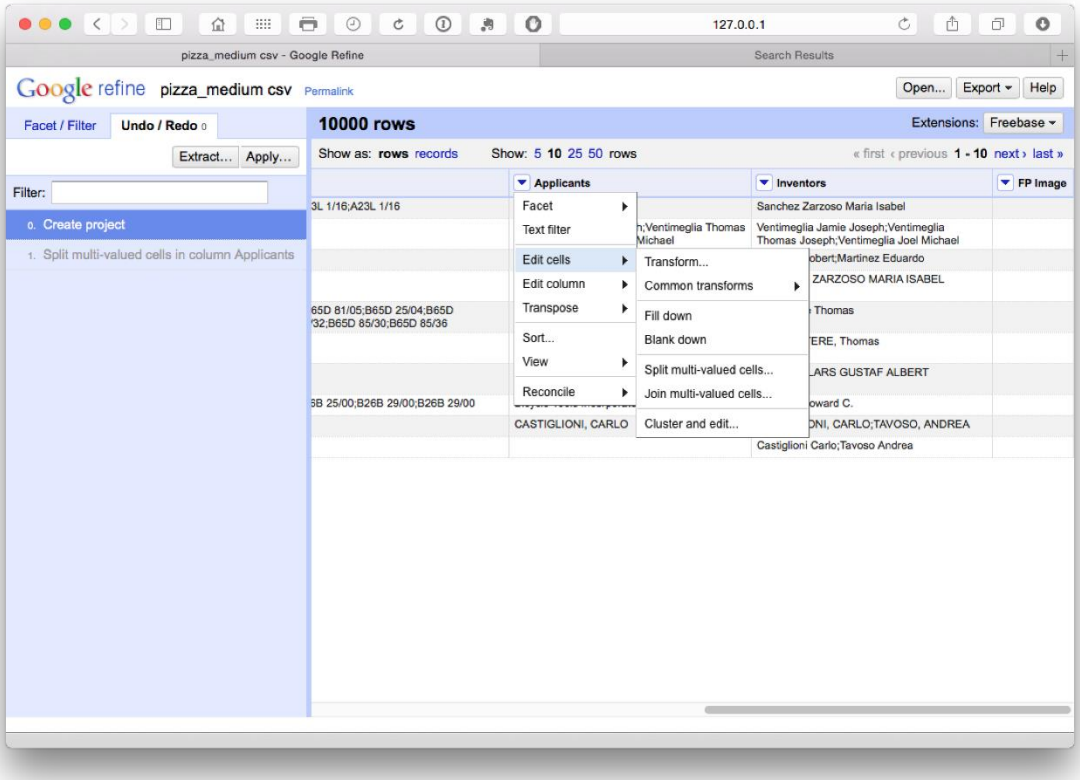


En este punto, podríamos comenzar el proceso de agrupación en clústeres para comenzar a limpiar los nombres que se discuten en la situación 2. Sin embargo, la desventaja de esto es que con este tamaño de conjunto de datos tendríamos que hacer esto 18 veces en ausencia de una manera fácil. de combinar las columnas en una sola columna (solicitantes) con un nombre en cada fila. Podríamos querer utilizar este enfoque en circunstancias en las que no nos centramos en los solicitantes y nos complace aceptar el primer nombre en la lista como el primer solicitante. En ese caso, simplemente estaríamos reduciendo el campo del solicitante a un solicitante. Tenga en cuenta que el primer solicitante que figura en la serie de nombres puede no ser siempre el primer solicitante que figura en una solicitud y no puede ser el nombre de una organización. Teniendo en cuenta estas advertencias, También podríamos usar este enfoque para reducir el campo de inventores concatenados a un inventor. Para fines generales que sería limpio y simple para fines de visualización.

Sin embargo, si quisiéramos realizar un análisis detallado del solicitante para un área de tecnología, tendríamos que adoptar un enfoque diferente.

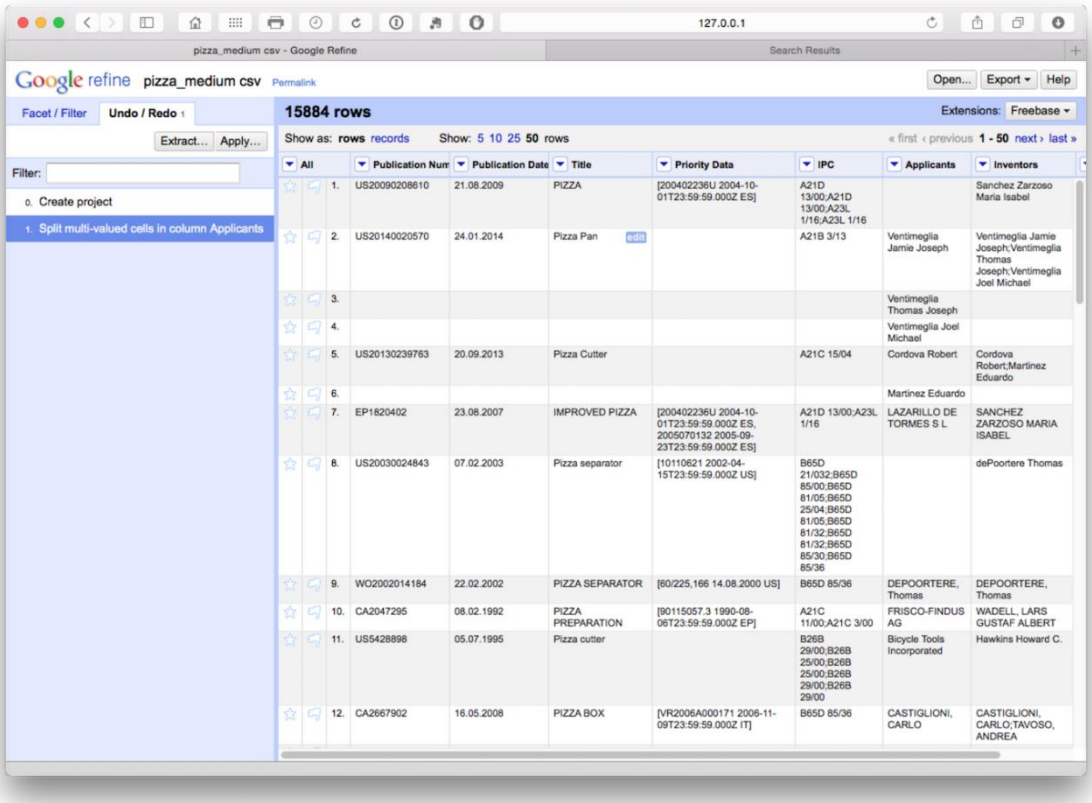
8.8.2 Situación 2 - Todos los solicitantes

Una de las fortalezas reales de Open Refine es que es muy fácil separar los nombres de solicitantes e inventores en filas individuales. En lugar de elegir Editar columna, ahora elegimos Edit cellsy luego split multi-valued cells.

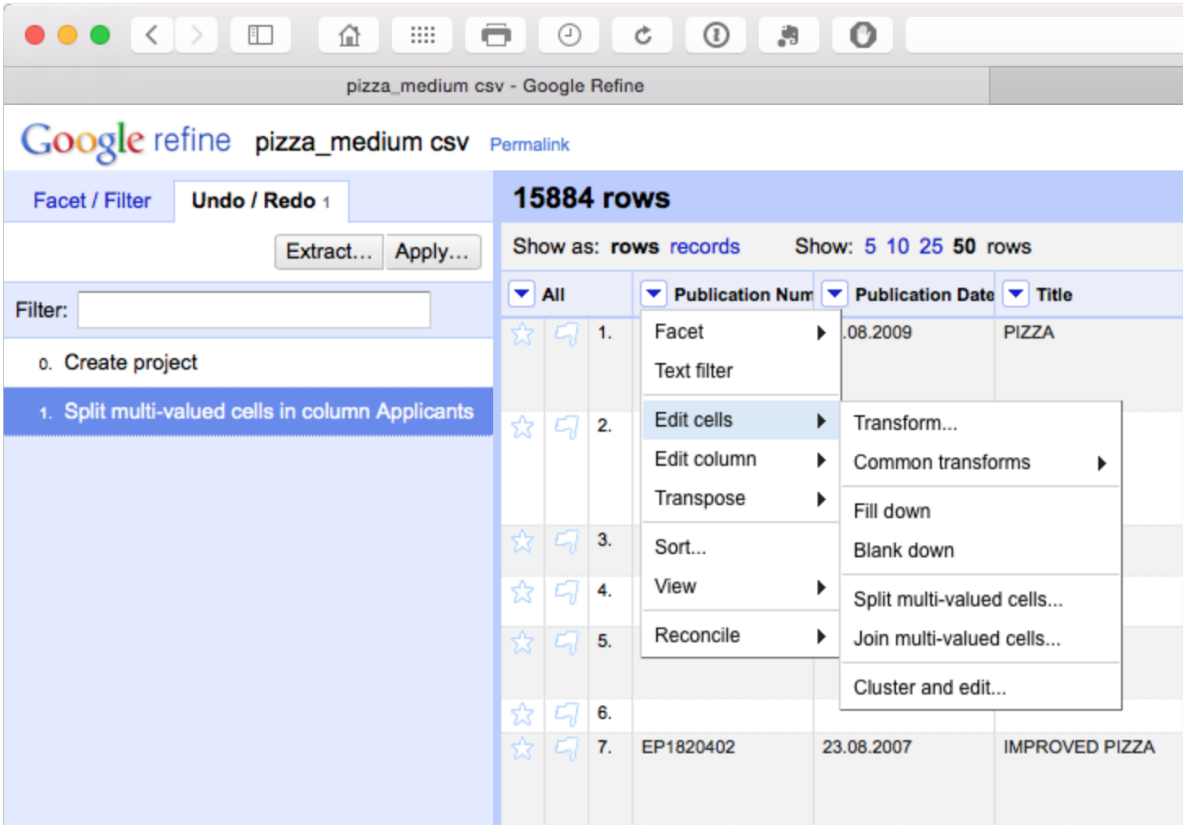


En el menú emergente, elija ;como separador en lugar de la coma predeterminada.

Ahora tenemos un conjunto de datos con 15,884 filas como podemos ver a continuación.

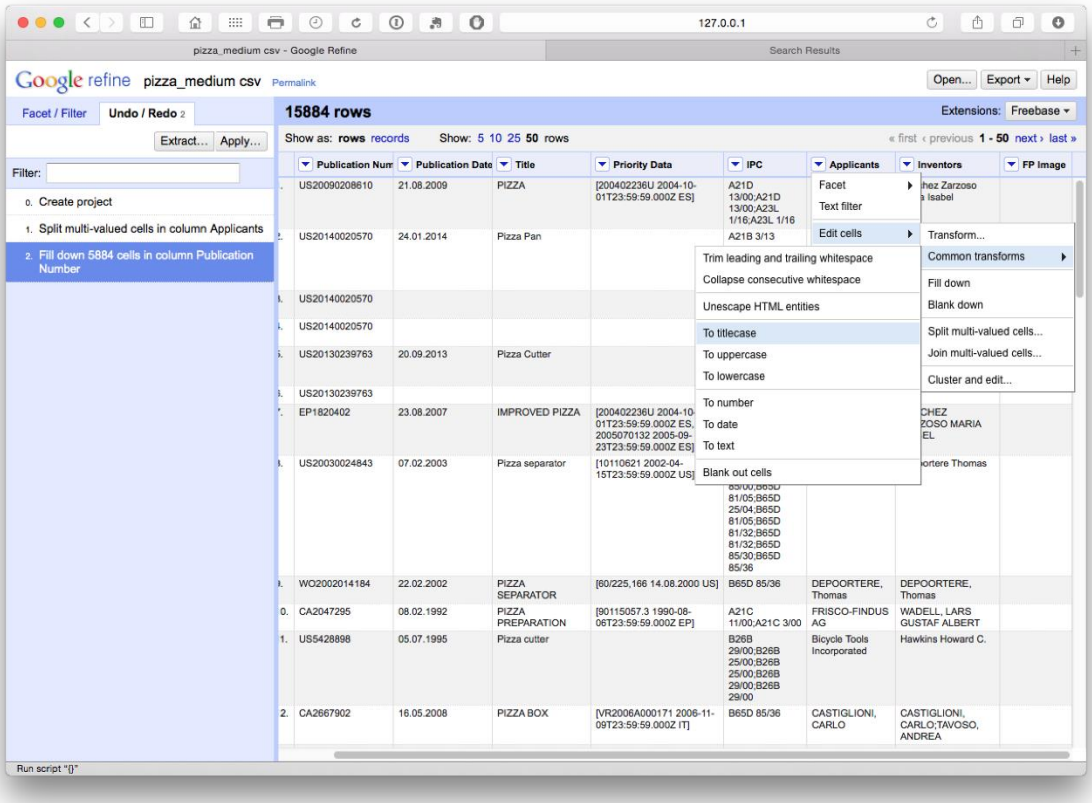


La ventaja de esto es que todos nuestros nombres individuales de solicitantes ahora están en una sola columna. Sin embargo, tenga en cuenta que el resto de los datos no se han copiado en las nuevas filas. Volveremos a esto, pero como precaución es sensato completar la columna del número de publicación como la clave que vincula a los solicitantes individuales con el registro. Así que hagámoslo por tranquilidad seleccionando publication number > edit cells > fill down.

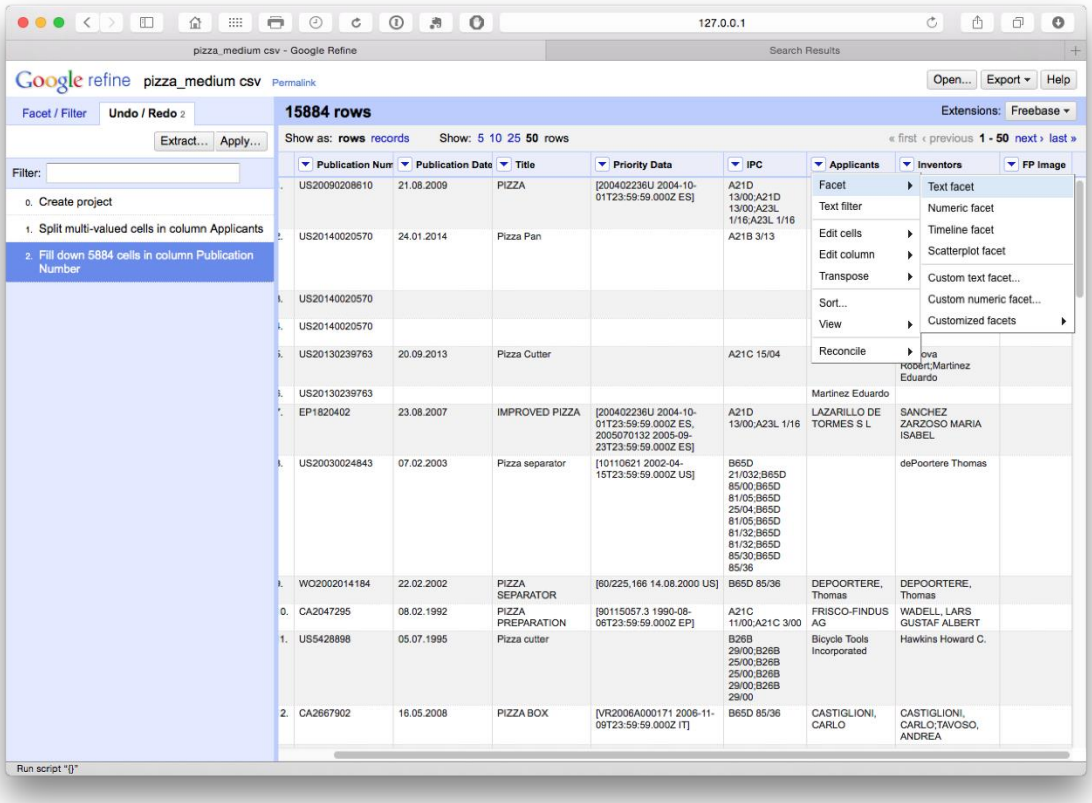


Ahora deberíamos tener una columna con los valores del número de publicación para cada solicitante como nuestra clave. Tenga en cuenta que es necesario tener cuidado al usar fill downOpen Refine como se explica en detalle aquí . Básicamente, el relleno no se realiza por registro, simplemente se llena. Eso puede significar que los datos se confunden. Esta es otra razón por la que es importante completar los valores en blanco con NA antes de comenzar a trabajar en Open Refine o como uno de los pasos iniciales de limpieza. El uso temprano de NA ayudará a evitar que el refinamiento llene las celdas en blanco con los valores de otro registro.

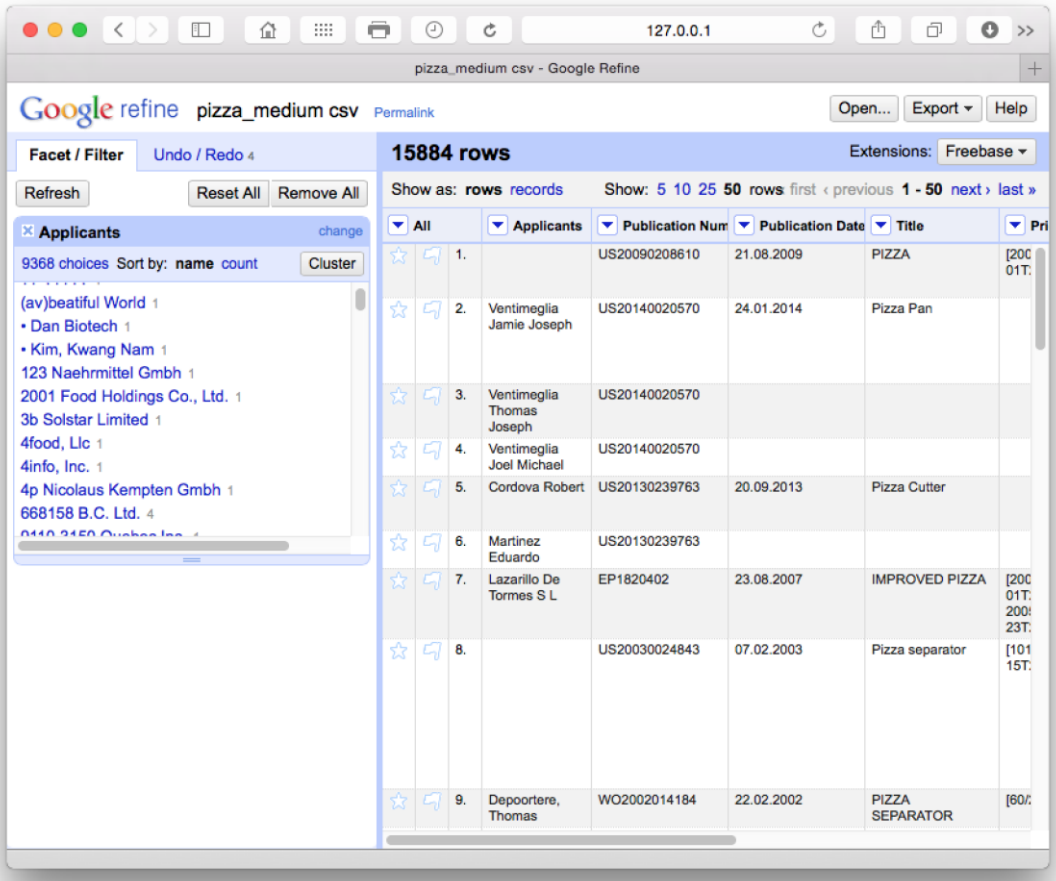
Si aún no lo ha hecho anteriormente para ayudar en el proceso de limpieza, y como buena práctica general, transforme el caso mixto en el campo de solicitantes a un solo tipo de caso. Para ello selecciona Applicants > Edit Cells > Common Transformations > To titlecase.



Ahora volvemos a los solicitantes y seleccionamos Facet > Text Facet.



Lo que veremos (con los solicitantes movidos a la primera columna al seleccionar Applicants > Edit Column > Move column to beginning) es una nueva ventana lateral con 9,368 opciones.



El botón de clúster activará un conjunto de seis algoritmos de limpieza con opciones de usuario en el camino. Vale la pena leer la documentación sobre estos pasos para decidir qué se ajustará mejor a sus necesidades en el futuro. Estos pasos de limpieza proceden de lo estricto a lo laxo en términos de criterios de coincidencia. El siguiente es un breve resumen de los detalles proporcionados en la página de documentación:

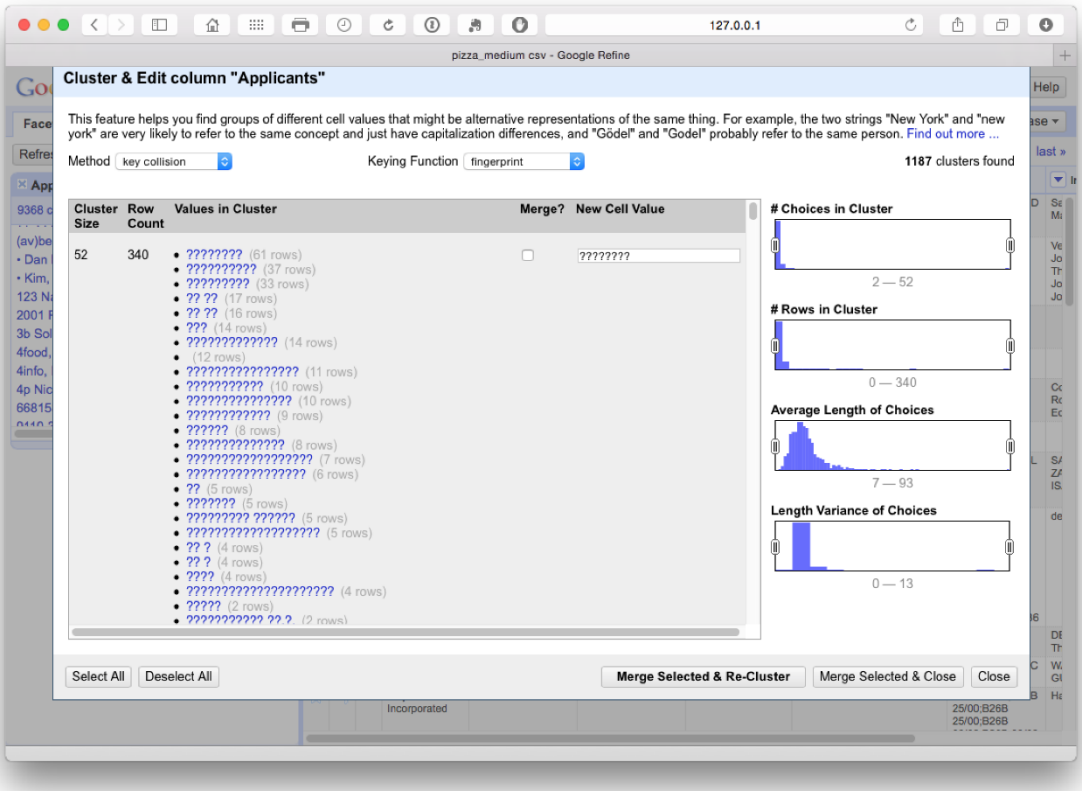
1. Huella dactilar Este método es el menos probable en el conjunto para producir falsos positivos (y eso es particularmente importante para los nombres de Asia oriental en los datos de patentes). Se trata de una serie de pasos que incluyen la eliminación de espacios en blanco finales, el uso de minúsculas, la eliminación de caracteres de puntuación y control, la división en tokens en el espacio en blanco, la división y unión y la normalización a ASCII.
2. Huella digital N-Gram. Esto es similar pero utiliza n-grams (una secuencia de caracteres o varias secuencias de caracteres) que se dividen, se ordenan,

se vuelven a unir y se normalizan a texto ASCII. La documentación destaca que esto puede producir más falsos positivos, pero es un alimento para encontrar grupos perdidos por las huellas dactilares.

3. Huella fonética. Esto transforma los tokens en la forma en que se pronuncian y produce diferentes huellas digitales en los métodos 2 y 3.
4. Métodos de vecinos más cercanos. Este es un método de distancia pero puede ser muy muy lento.
5. Levenshtein Distancia. Este famoso algoritmo mide el número mínimo de ediciones que se requieren para cambiar una cadena por otra (y por esta razón se conoce ampliamente como la distancia de edición). Por lo general, esto detectará errores tipológicos y ortográficos no detectados por los enfoques anteriores.
6. PPM es un uso particular de la complejidad de Kolmogorov como se describe en este artículo que se implementa en Open Refine as Prediction by Partial Matching.

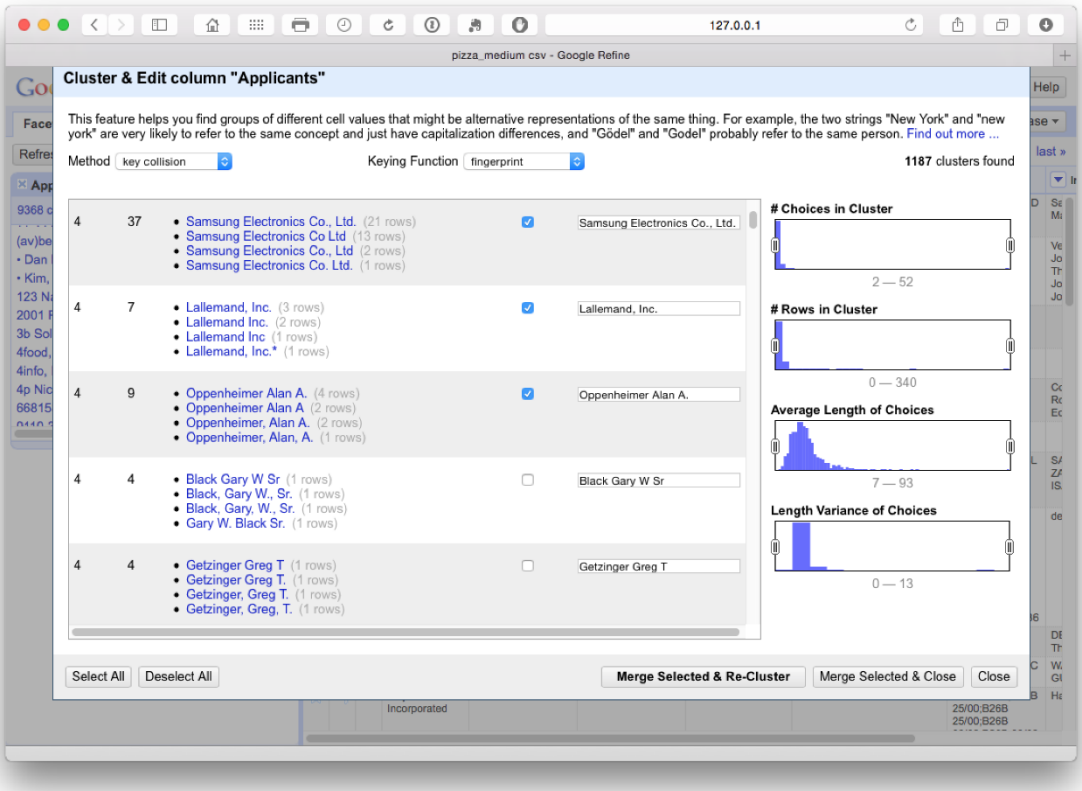
Es importante conocer estos métodos porque pueden afectar los resultados que recibe. En particular, se debe tener precaución con los nombres de Asia oriental, donde las tradiciones de nombres culturales producen una gran cantidad de falsos positivos en el mismo nombre para personas que en realidad son personas distintas (sinónimos o "agrupamiento" en la literatura). Esto puede tener impactos muy dramáticos en los resultados del análisis de patentes para los nombres de inventores porque tratará a todas las personas que comparten el nombre Smith, Johno Wang, Weicomo la misma persona, cuando en la práctica son múltiples personas individuales.

Ahora veremos cada algoritmo para ver los resultados.

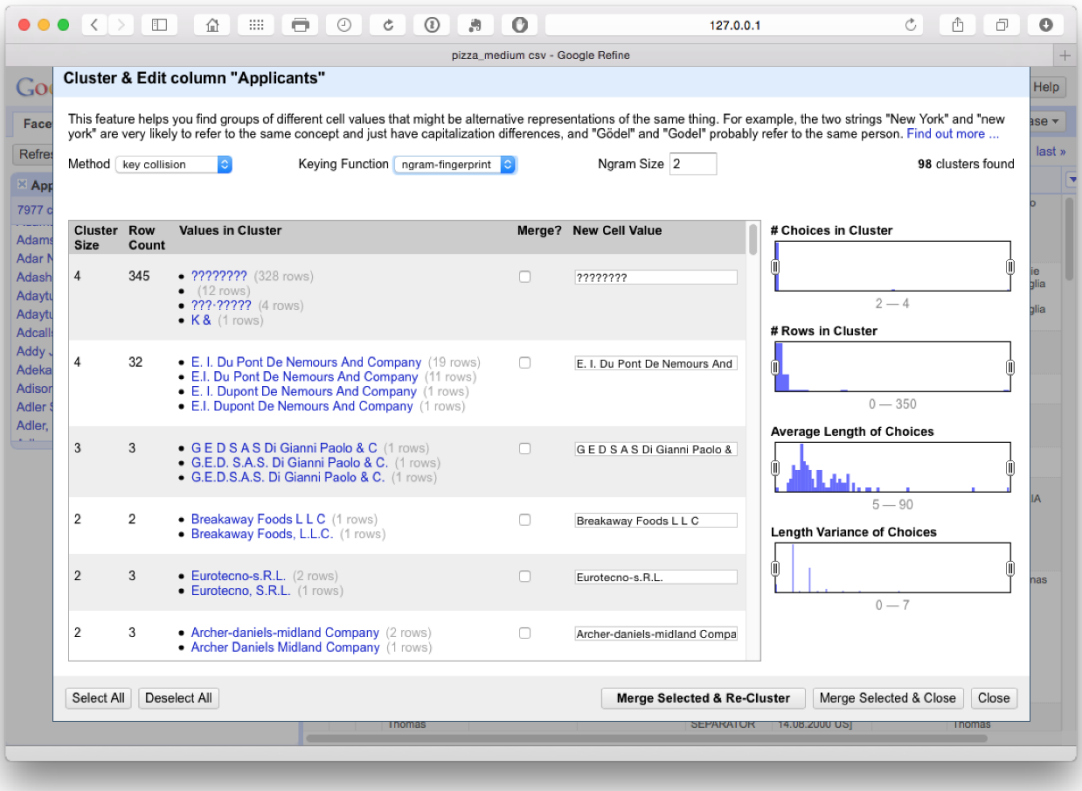


Esto identifica 1187 agrupaciones dominadas por caracteres ocultos en el campo de solicitantes (que suelen aparecer después del nombre). En esta etapa, debemos tomar algunas decisiones sobre si aceptar o rechazar la fusión propuesta marcando las Merge?casillas.

Este paso fue particularmente bueno para producir una coincidencia en nombres de variantes y reversiones de nombres como podemos ver aquí.

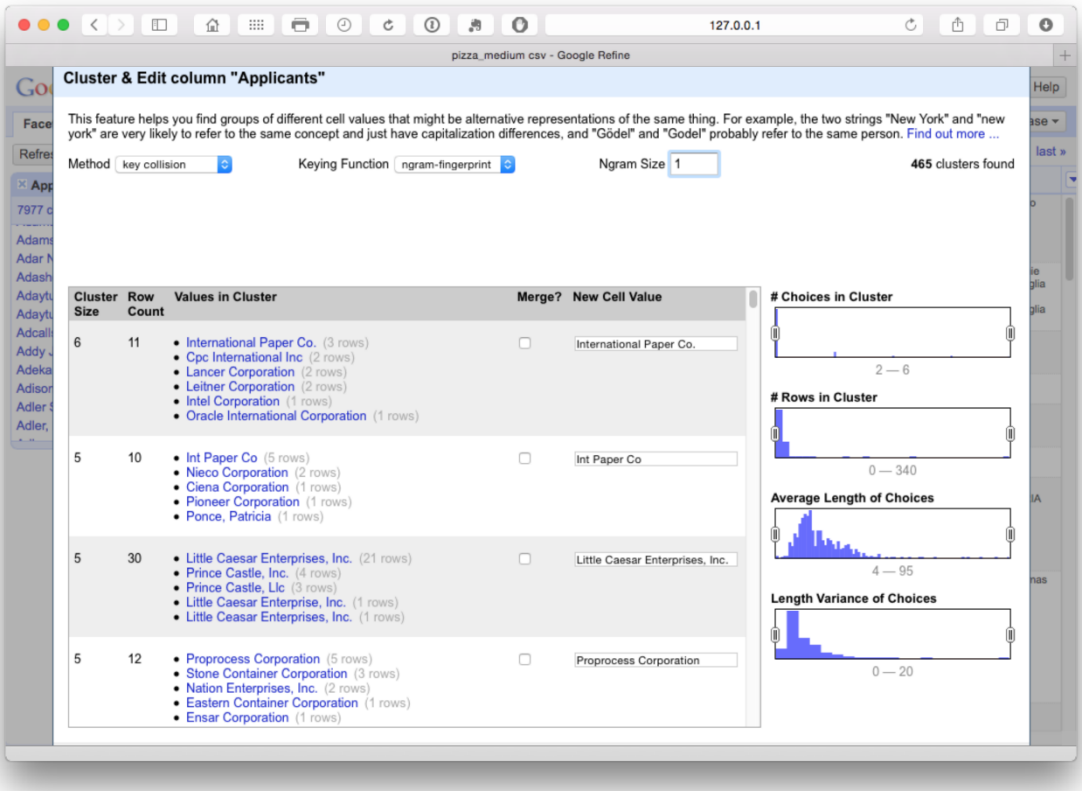


Vale la pena inspeccionar manualmente los datos antes de aceptarlos. Una opción importante aquí es usar el control deslizante Choices in Clusterpara mover el rango hacia arriba o hacia abajo y luego tomar una decisión sobre el punto de corte apropiado. Luego use seleccionar todo en la parte inferior izquierda para obtener resultados con los que esté satisfecho, seguido de Merge Selected & Re-Cluster. En el siguiente paso podemos cambiar el menú desplegable de la función de codificación a Ngram-fingerprint.



Esto produce 98 grupos que, según la inspección, son muy precisos. Los problemas a tener en cuenta aquí (y en todas partes) son nombres muy similares para compañías que pueden no ser la misma compañía (como Ltd. y Inc.) o divisiones distintas de la misma compañía. También es importante, al trabajar con nombres de inventores, no asumir que el mismo nombre es el mismo inventor en ausencia de otros criterios de coincidencia, o que las variaciones aparentemente menores en las iniciales (por ejemplo, Smith, John A y Smith, John B) son La misma persona porque bien pueden no serlo.

Para ver estos posibles problemas en acción, intente reducir el tamaño de N-gramos a 1. En este punto vemos lo siguiente.

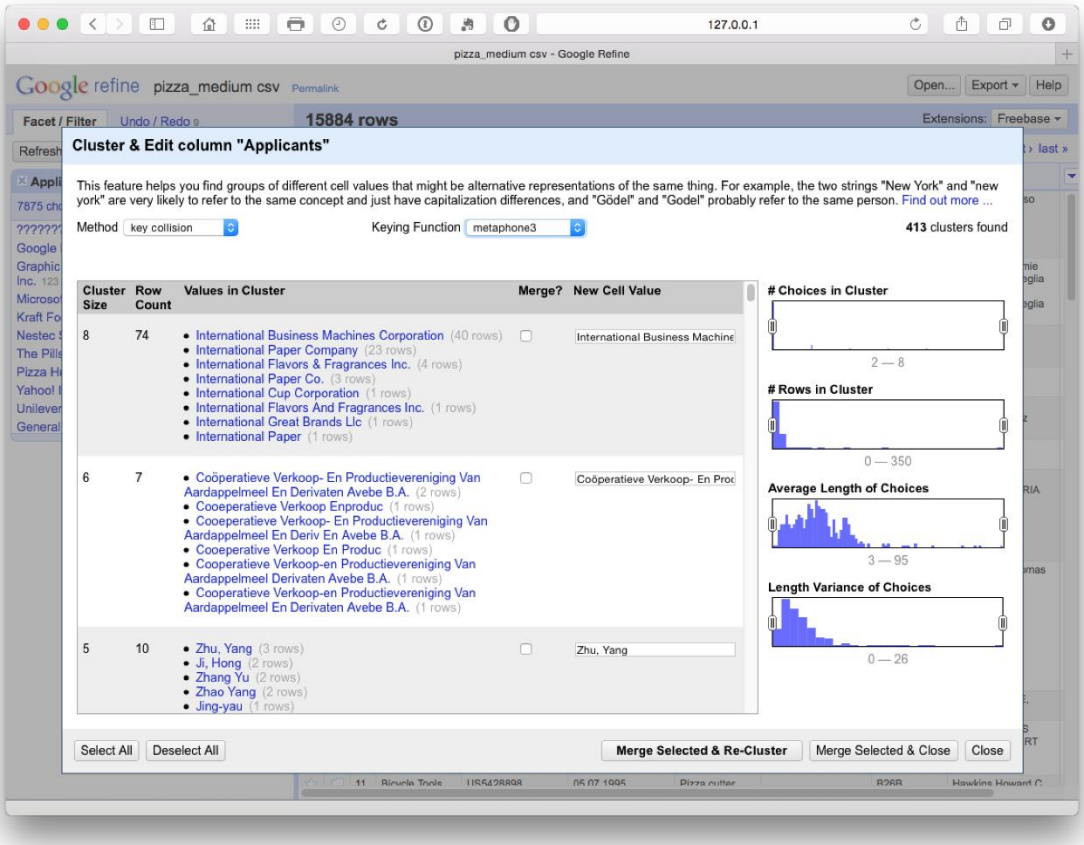


Esta medida es demasiado laxa y está agrupando compañías que no deben agruparse. En contraste, aumentar el valor de N-gramo a 3 o 4 apretará el grupo. Seleccionaremos todo en N-gram 2 y procederemos al siguiente paso.

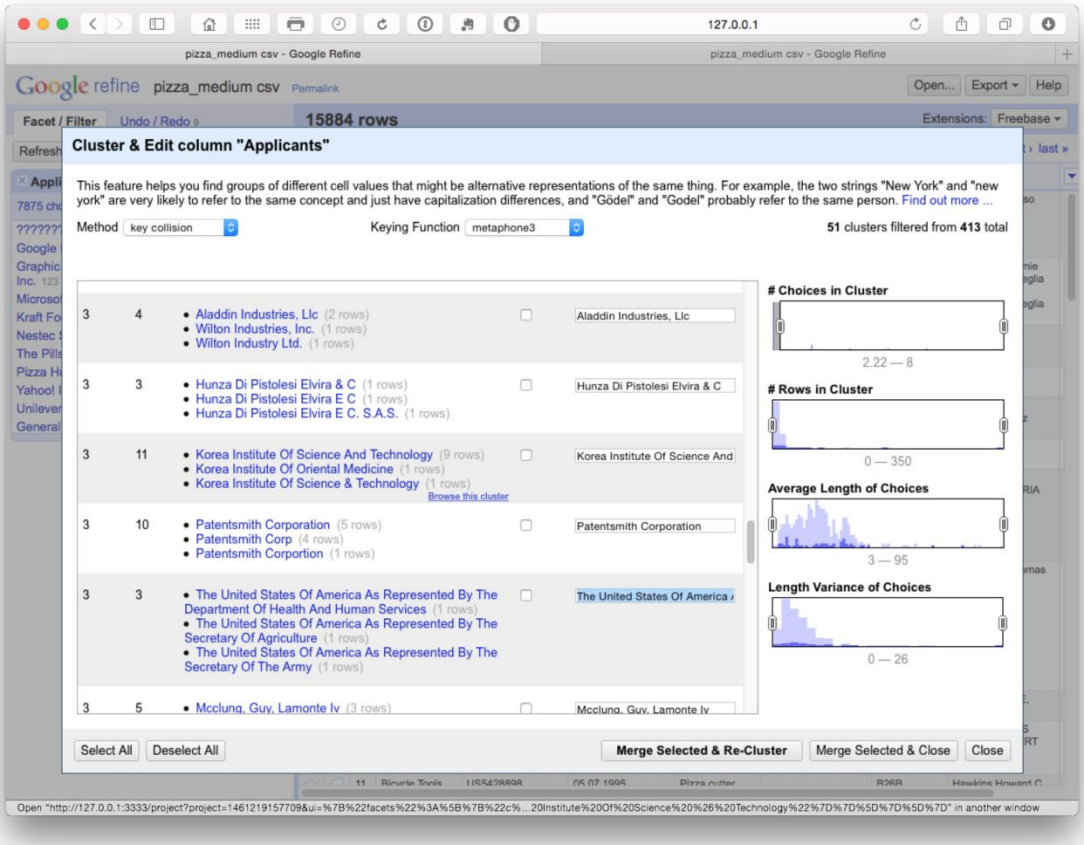
En este punto, vale la pena señalar que los 9,368 clusters originales se han reducido a 7,875 y si clasificamos el conteo en la ventana principal, entonces Google está comenzando a emerger como el principal candidato en nuestro conjunto de 10,000 registros.

8.8.3 Agrupación de huellas dactilares fonéticas (Metaphone 3)

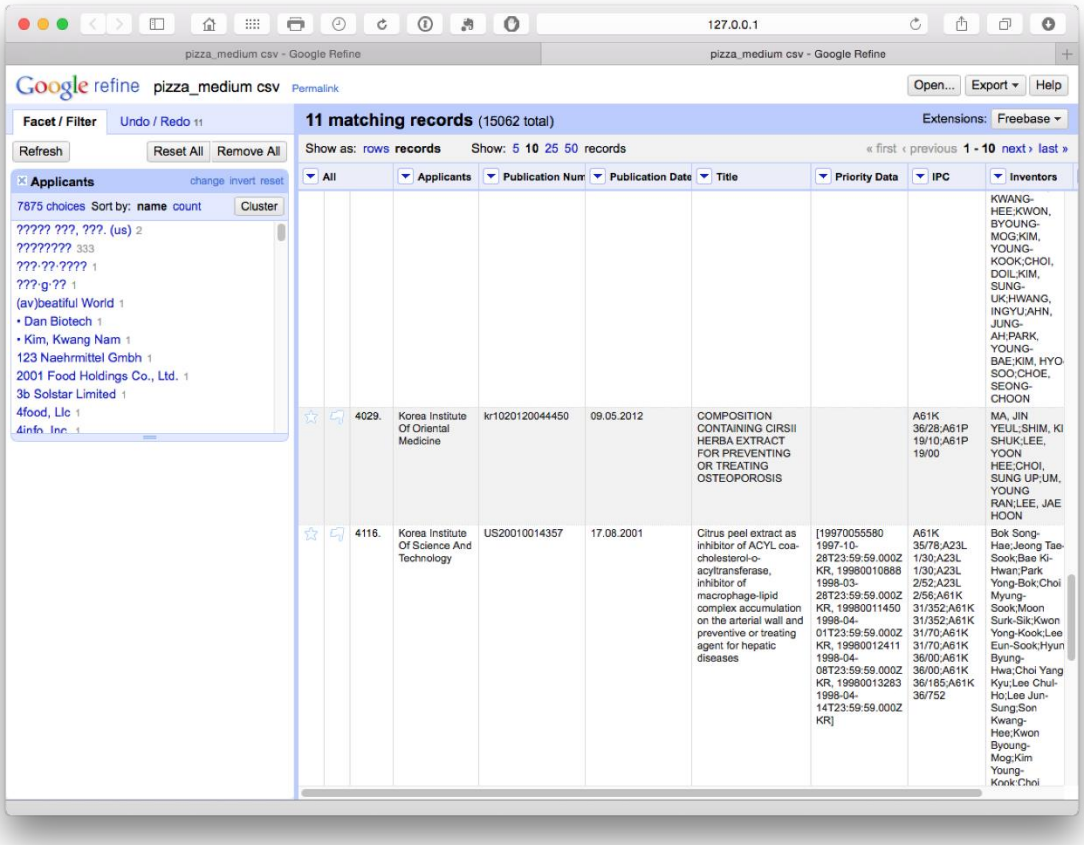
Como podemos ver a continuación, la agrupación de Metaphone 3 produce 413 clústeres más sueltos con falsos positivos en International Business Machines, pero positivos en Cooperativa Verkoop.



En este punto, podríamos revisar manualmente los 413 clústeres y seleccionarlos según corresponda o cambiar la configuración para reducir el número de clústeres utilizando el Choices in Clusterscontrol deslizante hasta que veamos algo manejable para la revisión manual. En esta etapa también podríamos usar la browse this clusterfunción que aparece al pasar el mouse sobre una selección en particular, para revisar los datos (consulte la segunda entrada en la imagen a continuación).

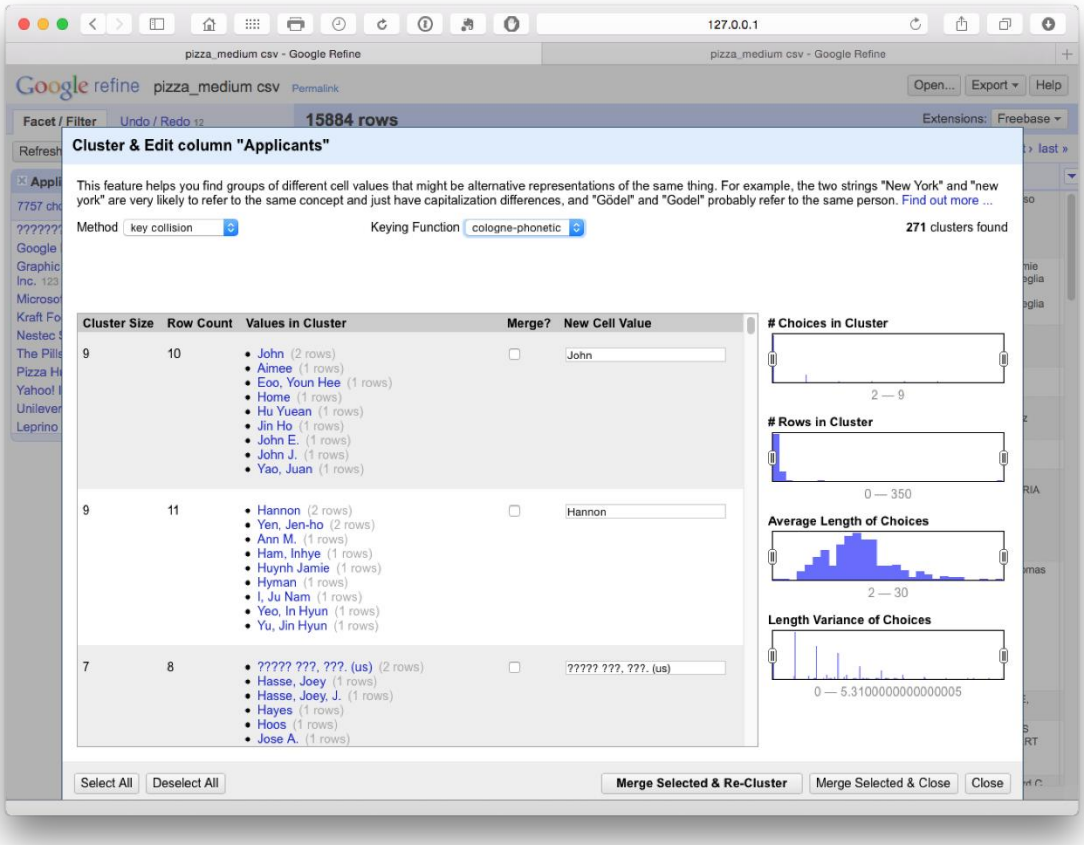


En este caso, estamos tratando de determinar si el Instituto de Medicina Oriental de Corea se debe agrupar con el Instituto de Ciencia y Tecnología de Corea (lo que parece poco probable). Si abrimos la función del navegador, podemos revisar las entradas para características compartidas como posibles criterios de coincidencia.



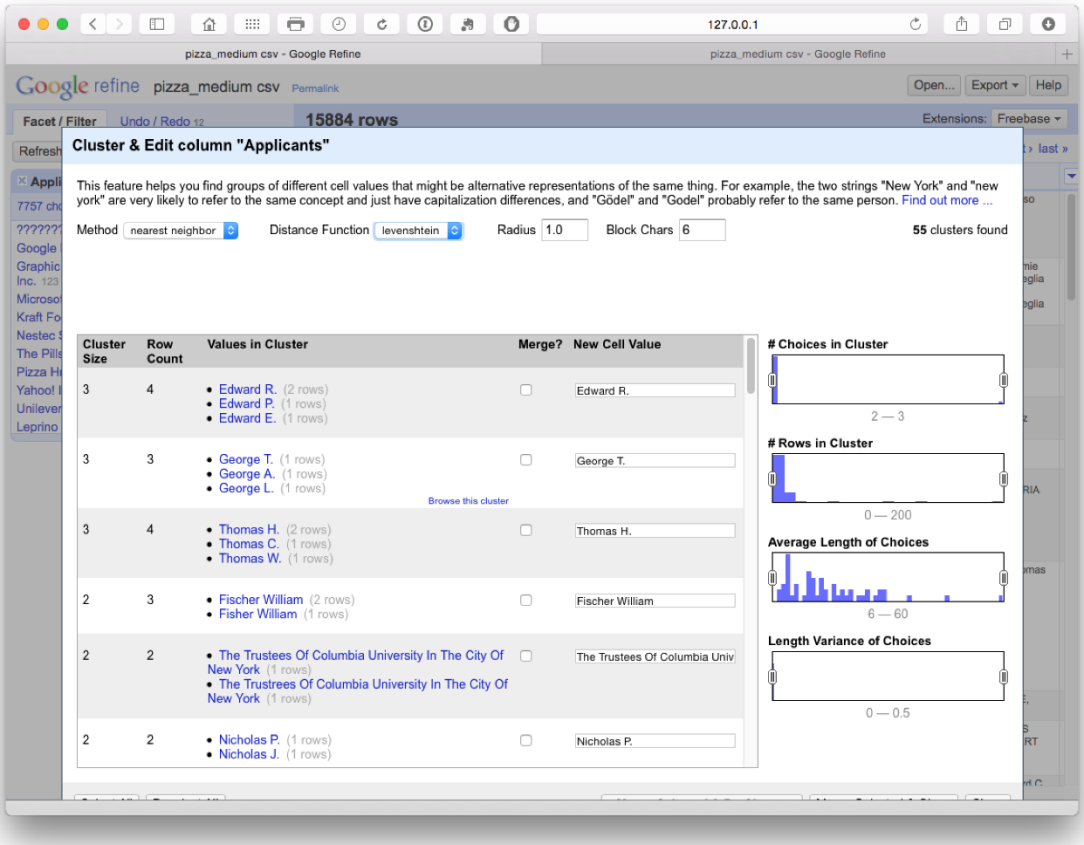
Por ejemplo, si los solicitantes compartieron inventores y / o el mismo título, es posible que deseamos registrar este registro en la agrupación más grande (recordando que hemos exportado los datos originales limpiados). O, como es más probable, podríamos capturar a la mayoría de los miembros del grupo y eliminar el Instituto de Medicina Oriental de Corea. Sin embargo, cómo hacer esto no es del todo obvio.

En la práctica, la selección de elementos en esta etapa se alimenta a la siguiente etapa de limpieza utilizando el algoritmo fonético de Colonia. Como podemos ver a continuación, este algoritmo identificó 271 agrupaciones que estaban agrupadas casi en su totalidad en los nombres de individuos con un número limitado de resultados precisos.



8.8.4 Levenshtein Editar distancia

Los pasos finales en el proceso se centran en las coincidencias del vecino más cercano para nuestro reducido número de agrupaciones. Tenga en cuenta que esto puede tardar un poco en ejecutarse (por ejemplo, 10-15 minutos para los grupos de +7,000 en este caso). Los resultados se muestran a continuación.



En algunos casos, la configuración predeterminada coincidía con los nombres individuales en diferentes iniciales, pero en la mayoría de los casos los grupos parecían válidos y se aceptaron. En este caso, se requiere especial precaución con los nombres de las personas y la navegación en los resultados para verificar la precisión de las coincidencias.

8.8.5 PPM

El paso PPM es el logaritmo final, pero tomó tanto tiempo que decidimos abandonarlo en relación con las ganancias probables.

8.8.6 Preparándose para la exportación

En la práctica, el proceso de limpieza generará una nueva tabla de datos para la exportación que se centrará en las características de los solicitantes. Para preparar la exportación con el grupo limpio de nombres de solicitantes, habrá una opción sobre si retener el número de publicación como clave única o si usar el proceso de relleno ilustrado arriba en las columnas del conjunto de datos. Es importante tener en cuenta que se requiere precaución en el uso general del relleno.

También es importante tener en cuenta que el conjunto de datos que se ha creado contiene muchas más filas que la versión original. Antes de exportar sugerimos dos pasos:

1. Vuelva a ejecutar Transformaciones comunes> en el caso del título, para regularizar los nombres que puedan haberse omitido en la primera ronda y volver a ejecutar los espacios en blanco de recorte para cualquier espacio que surja de la división de nombres.
2. Vuelva a ejecutar las facetas en cada columna, seleccione el espacio en blanco al final del panel de facetas y rellene con NA. Alternativamente, haga esto inmediatamente después de la exportación.

8.9 Round Up

En este capítulo hemos cubierto las características principales de la limpieza básica de datos utilizando Open Refine. Como debería quedar claro, aunque requiere una inversión en la familiarización, es una herramienta poderosa para limpiar conjuntos de datos de patentes de tamaño pequeño a mediano, como nuestros registros de patentes de 10,000 pizzas. Sin embargo, se requiere un grado de paciencia, precaución y planificación anticipada para crear un flujo de trabajo efectivo con esta herramienta. Es probable que una mayor inversión de tiempo (como el uso de expresiones regulares en GREL) mejore las tareas de limpieza antes del análisis.

Open Refine también es probablemente la herramienta gratuita más fácil de usar para separar y limpiar nombres de solicitantes e inventores sin conocimientos de programación. Solo por esa razón, mientras observa las advertencias resaltadas anteriormente, Open Refine es una herramienta muy valiosa en la caja de herramientas de análisis de patentes de código abierto.

8.10 Recursos útiles

El sitio web de Open Refine tiene enlaces a muchos recursos útiles que incluyen tutoriales en video

Abrir Refinar Recetas Wiki

Abrir Refinar consejos y trucos

Preguntas de desbordamiento de pila en Open Refine

Capítulo 9 Tableau Public

9.1 Introducción

En este capítulo analizaremos y visualizaremos los datos de patentes usando Tableau Public.

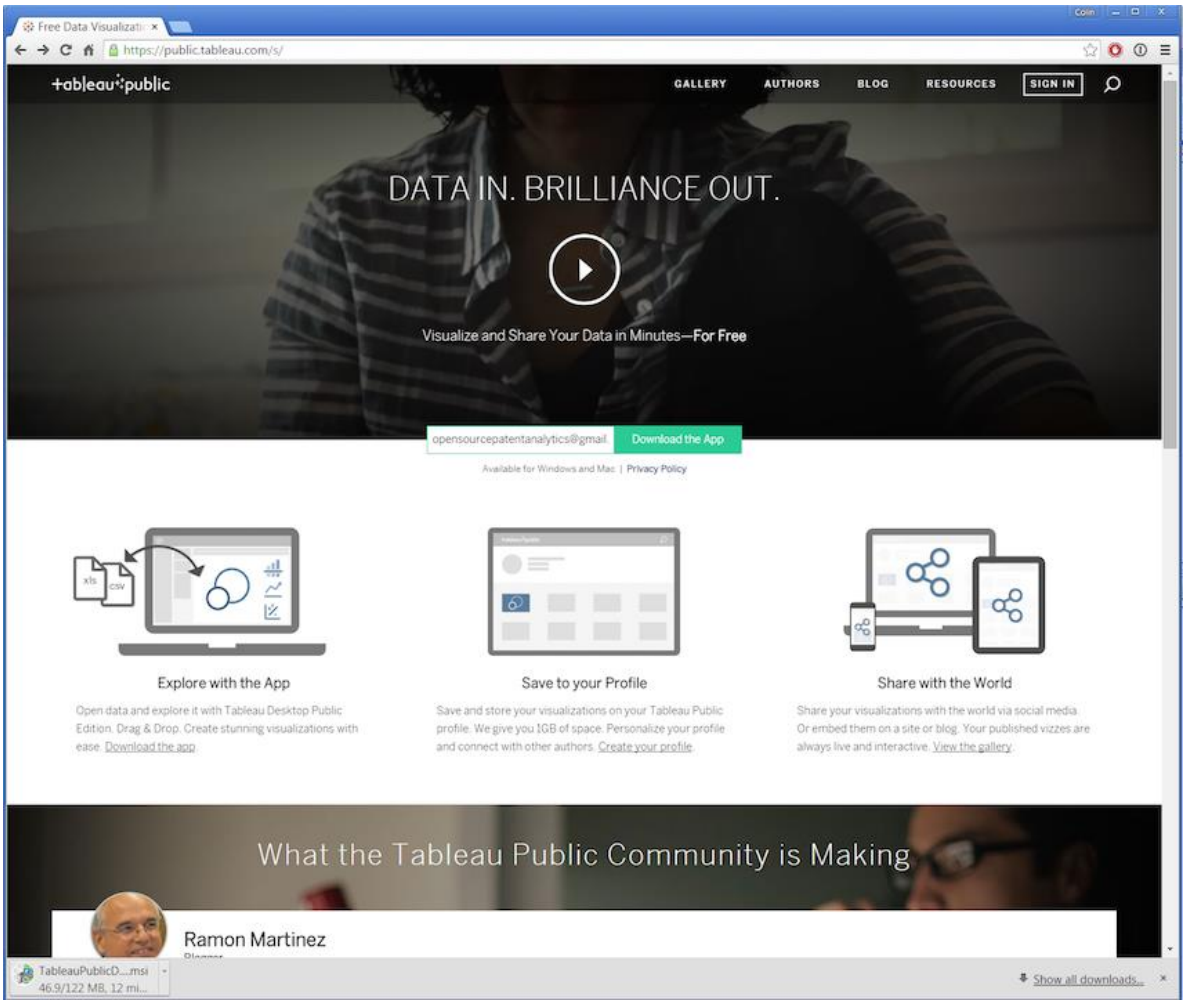
Tableau Public es una versión gratuita de Tableau Desktop y proporciona una muy buena introducción práctica al uso de datos de patentes para análisis y visualización. En muchos casos, Tableau Public representará el estándar que otras herramientas de código abierto y gratuitas deberán cumplir.

Esta es una demostración práctica del uso de Tableau en el análisis de patentes. Hemos creado un conjunto de tablas de datos de patentes limpias al pizza patentsusar una muestra de 10,000 registros de WIPO Patentscope que puede descargar como un archivo .zip desde aquí para usar durante el tutorial. Los detalles del proceso de limpieza para llegar a esta etapa se proporcionan en el libro de códigos que se puede ver aquí . El tutorial Open Refine se puede usar para generar archivos limpios muy similares a los utilizados en este tutorial utilizando sus propios datos. No necesitará limpiar ningún dato utilizando nuestros archivos de conjunto de entrenamiento.

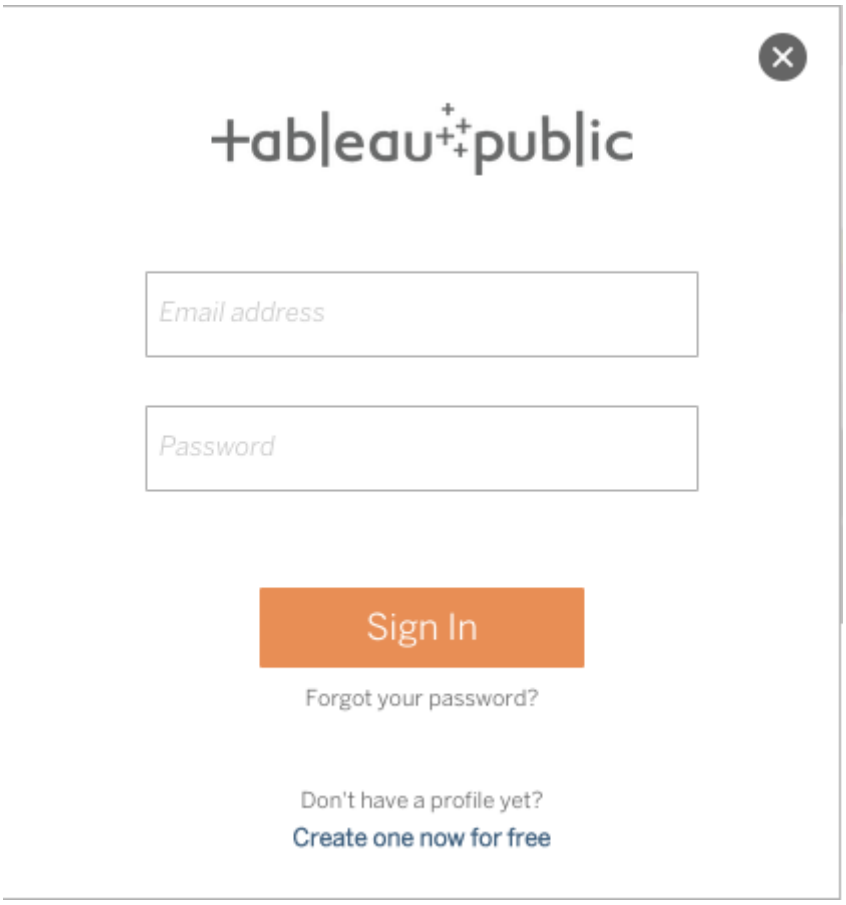
Este artículo lo guiará a través de las características principales de Tableau Public y los tipos de análisis y visualización que se pueden realizar utilizando Tableau. En el proceso, creará algo muy similar a este libro de trabajo .

9.2 Instalación de Tableau

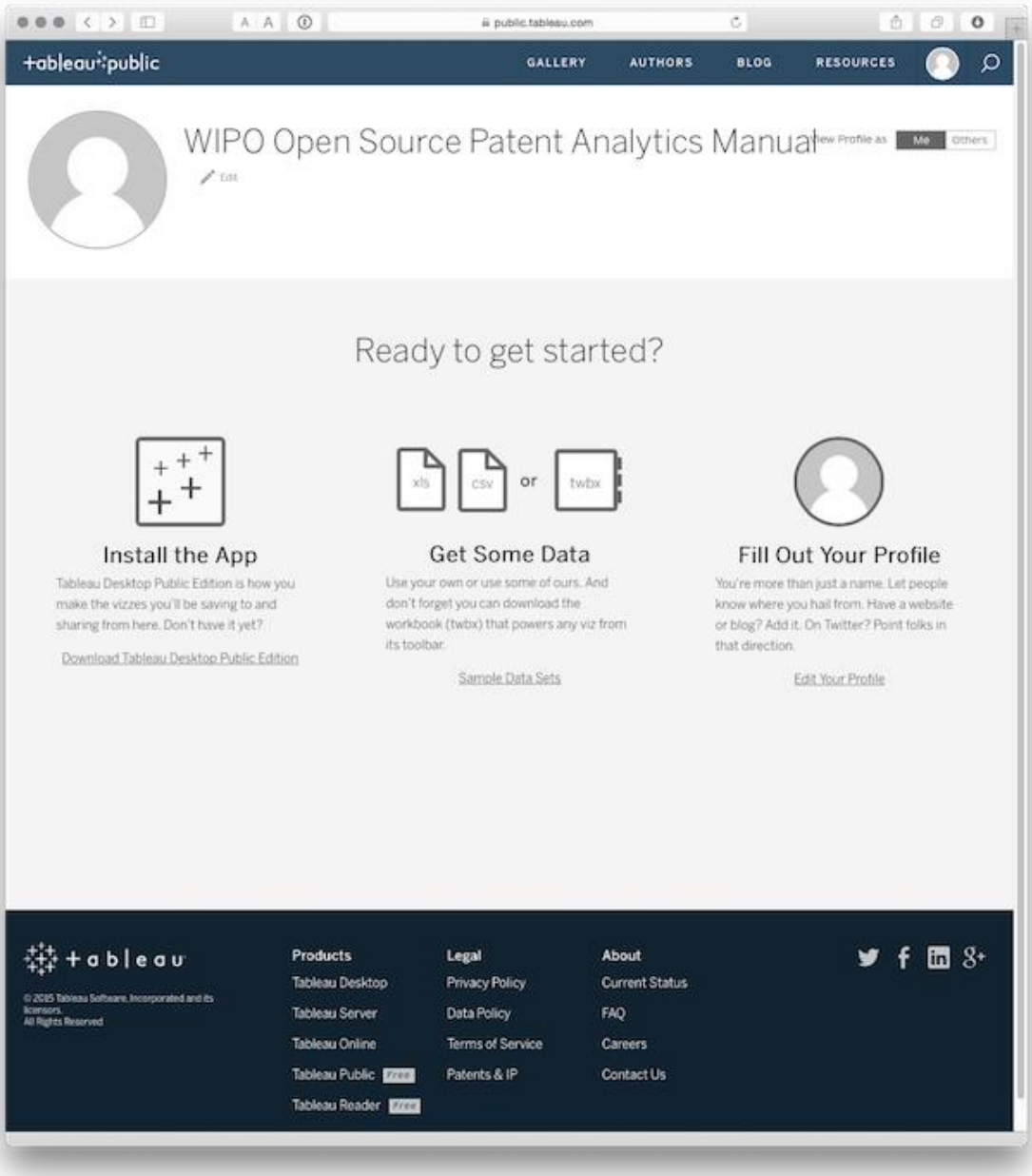
Tableau se puede instalar para su sistema operativo visitando el sitio web de Tableau Public e ingresando su dirección de correo electrónico como se muestra en la imagen a continuación.



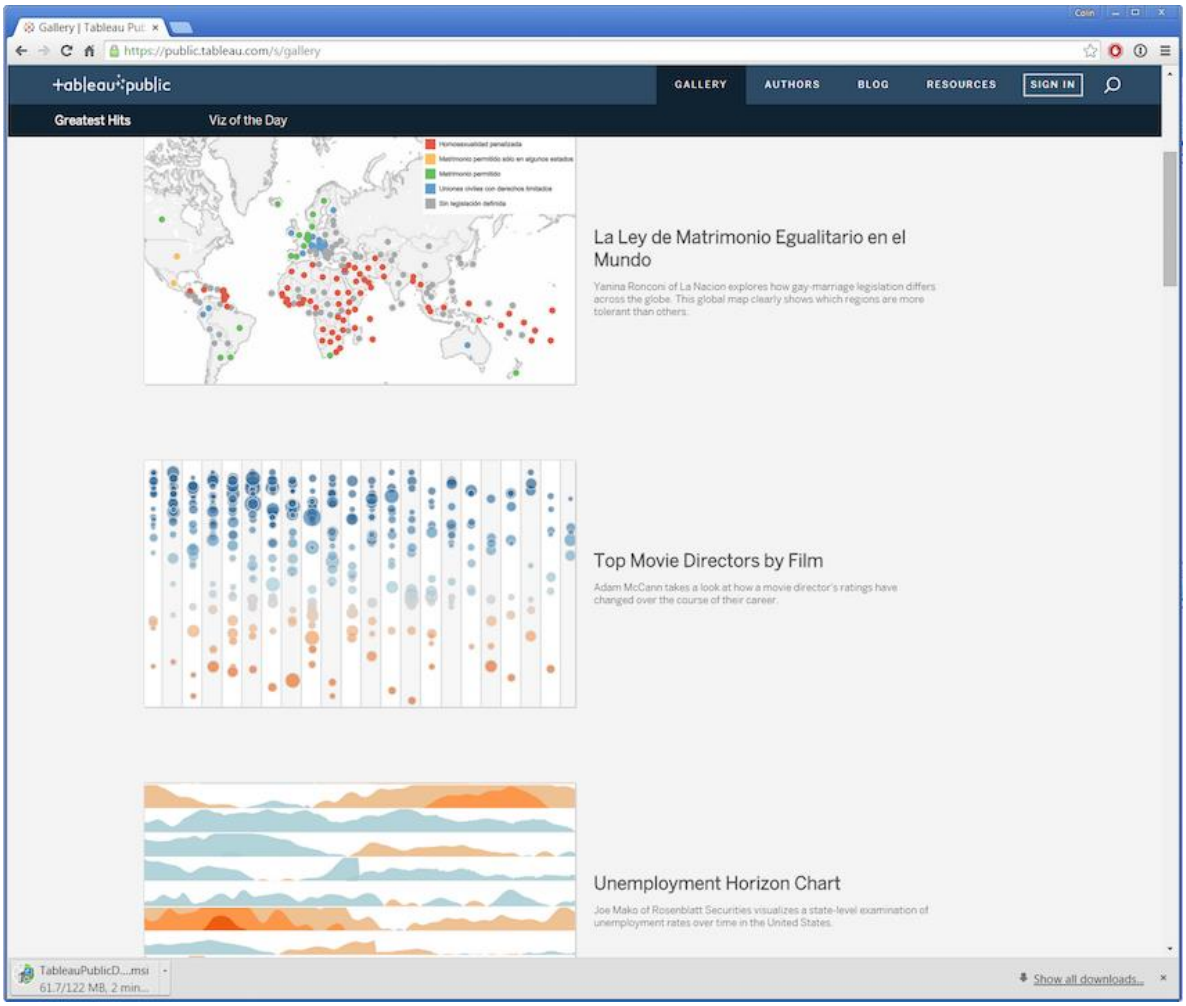
Mientras espera la descarga de la aplicación, es una buena idea seleccionar Sign Iny luego Create one now for Freeregistrarse para obtener una cuenta pública de Tableau que le permita cargar sus libros de trabajo en la web y compartirlos. Abordaremos los problemas de privacidad al hacer que los libros de trabajo sean públicos o privados a continuación, pero como su nombre indica, Tableau Public no es para información comercial confidencial.



Esto te llevará a una página de perfil vacía.

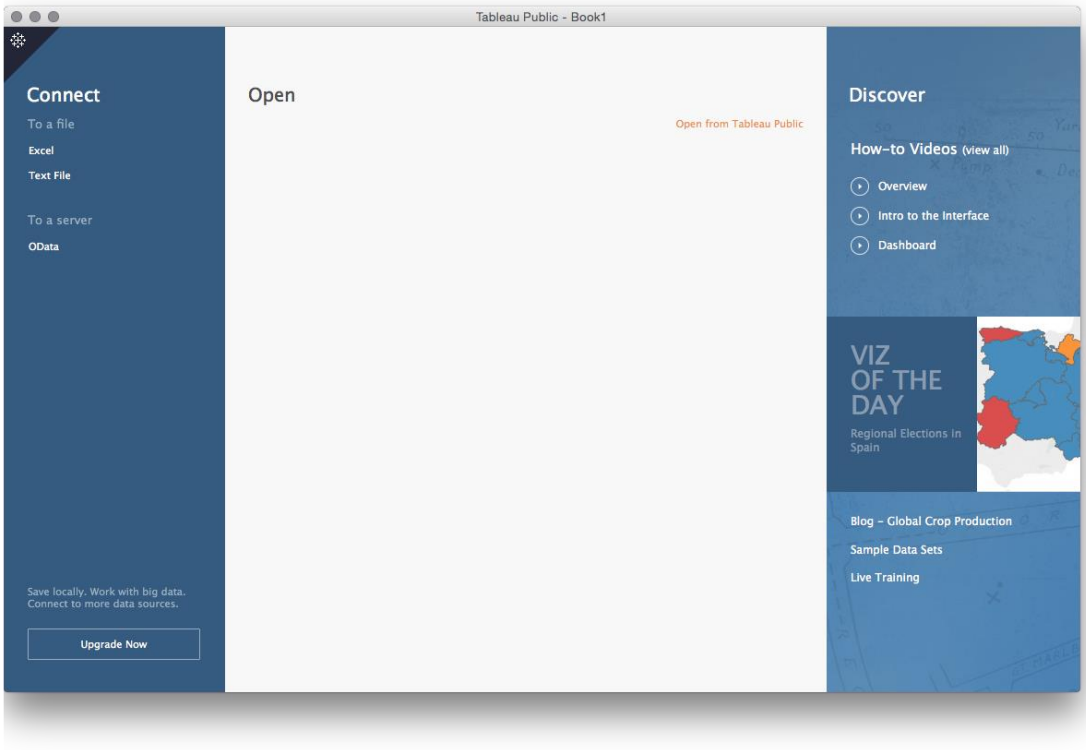


Mientras esté allí, es posible que desee consultar la Galería de otros libros de trabajo de Tableau Public para obtener algunas ideas sobre lo que es posible lograr con Tableau. Es posible que desee ver un Libro de trabajo de Tableau para la literatura científica que acompaña a este artículo de PLOS ONE sobre biología sintética . Si bien ahora tiene algunos años, da una idea de las posibilidades de Tableau y la sensación de una página de perfil existente.

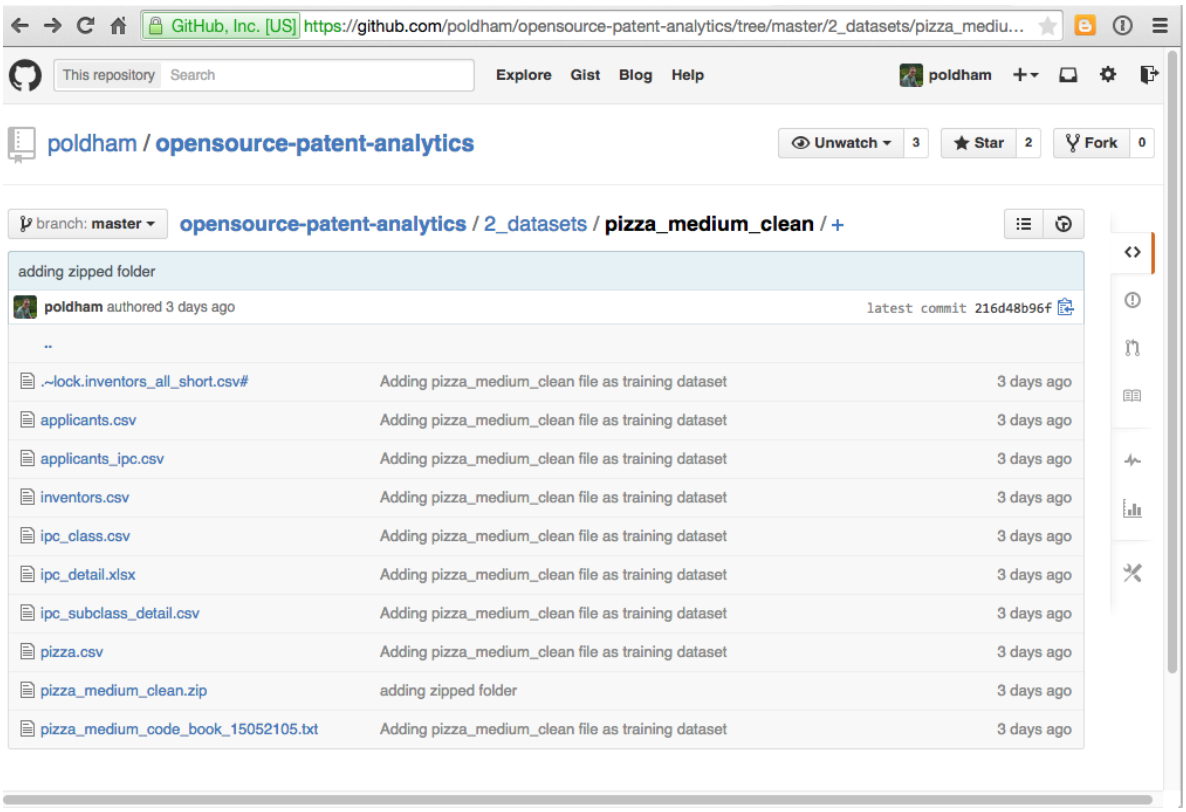


9.3 Cómo empezar

Cuando abres la aplicación por primera vez, verás una página en blanco. Antes de cargar algunos datos, tenga en cuenta lo útil How-to-Videosa la derecha y el enlace a a visualisation of the day. También hay bastantes videos de capacitación aquí y un foro comunitario muy útil . Si te quedas atascado, o te preguntas cómo alguien produjo una visualización genial, este es el lugar para ir.

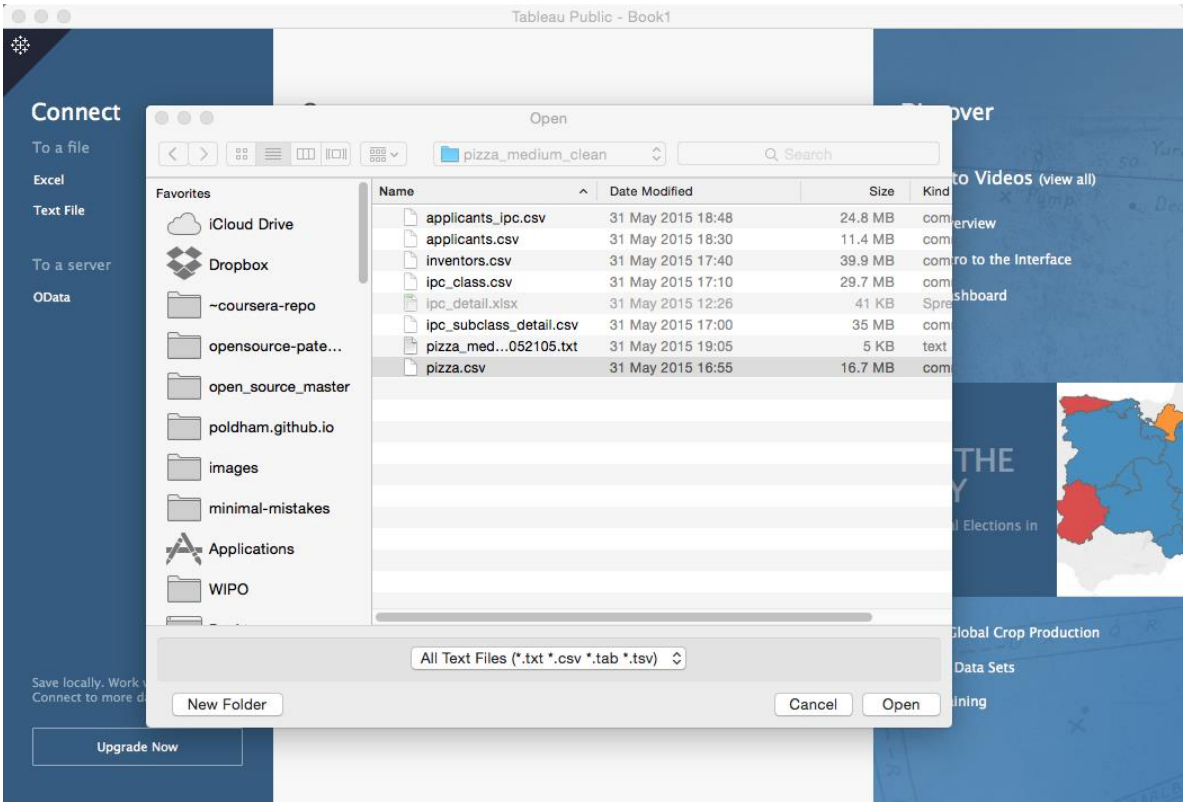


Para evitar mirar una página en blanco, ahora necesitamos cargar algunos datos. En Tableau Public, esto está limitado a archivos de texto o Excel. Para descargar los datos como un solo .ziparchivo, haga clic aquí o visite el repositorio de GitHub . descomprima el archivo y verá una colección de archivos .csv. El archivo de Excel y el libro de códigos deben ignorarse como complementarios.

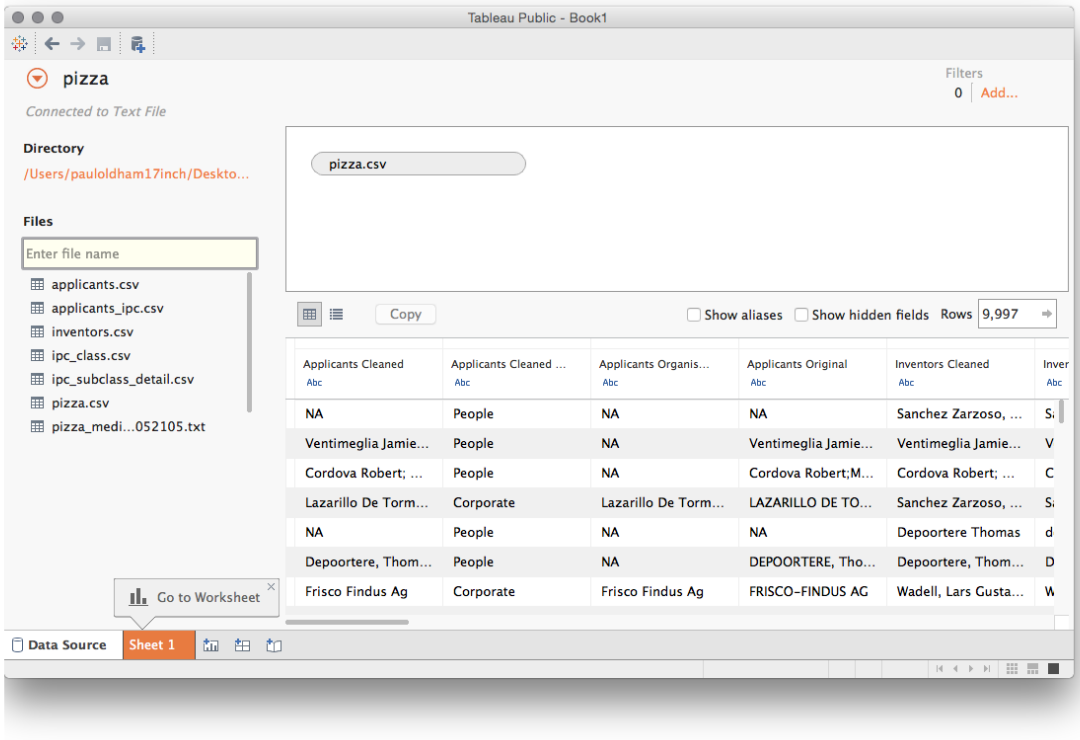


Como podemos ver arriba, hay una serie de archivos en este conjunto de datos. El core archivo o referencia es pizza .csv. Todos los demás archivos son aspectos de ese archivo, como solicitantes, inventores y códigos de clasificación de patentes internacionales. Esos campos concatenados en pizza han sido separados y limpiados. Un archivo, applicants_ipces un archivo secundario applicants que nos permitirá acceder a información de IPC para solicitantes individuales. Puede que esto no tenga mucho sentido en este momento, pero no se preocupe, lo hará en breve.

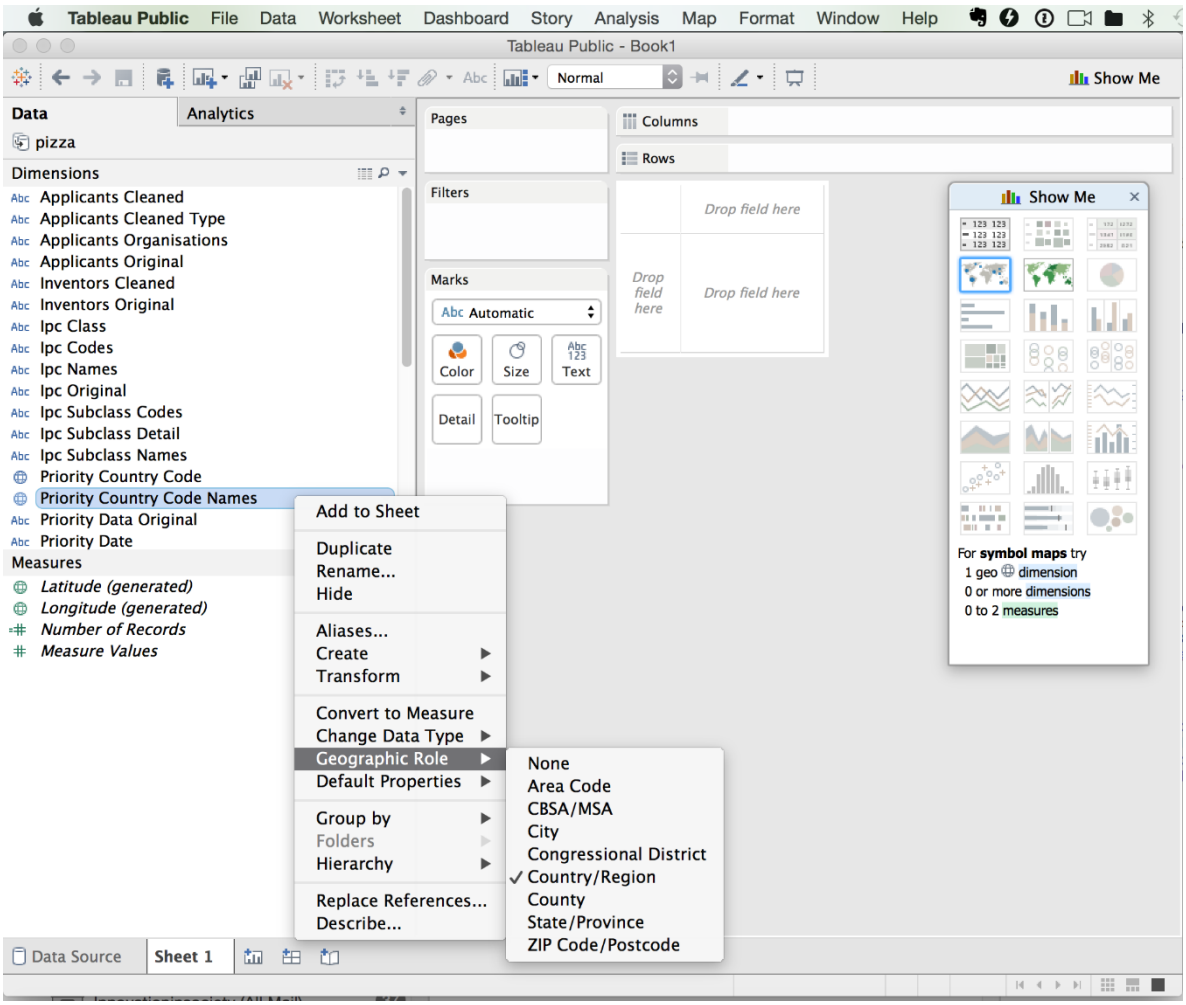
Para empezar seleccionaremos el pizza archivo .csv:



Luego veremos una nueva pantalla que muestra algunos de los datos y los otros archivos en la carpeta. En la parte inferior hay una bandera con Go to Worksheet, así que vamos a hacer eso.



Ahora veremos una pantalla que se divide en a Dimensionsla izquierda, con la Measuressiguiente. Podemos ver que en las dimensiones hay una gran cantidad de campos de datos. Tenga en cuenta que Tableau intentará adivinar el tipo de datos (por ejemplo, la información numérica o de fecha está marcada con #, los datos geográficos están marcados con un globo terráqueo, los campos de texto están marcados con Abc). Tenga en cuenta que Tableau no siempre tiene este derecho y que es posible cambiar un tipo de datos seleccionando un campo y haciendo clic derecho como podemos ver a continuación.



En el lado derecho podemos ver un menú de panel flotante. Esto se puede ocultar como una barra de menú haciendo clic en la x. Este panel muestra las opciones de visualización que están disponibles para el campo de datos que hemos seleccionado. En este caso, hay dos opciones de mapas disponibles porque Tableau ha reconocido automáticamente los nombres de los países como información geográfica. Tenga en cuenta que persuadir a Tableau para que presente la opción que desea (por ejemplo, visualizar datos de año en año como un gráfico de líneas) puede implicar cambiar la configuración del campo hasta que la opción que desea esté disponible.

En la parte inferior de la pantalla veremos un número de hoja de trabajo Sheet 1 y luego opciones para agregar tres tipos de hojas:

1. Una nueva hoja de trabajo
2. Un nuevo tablero de mandos
3. Una nueva historia

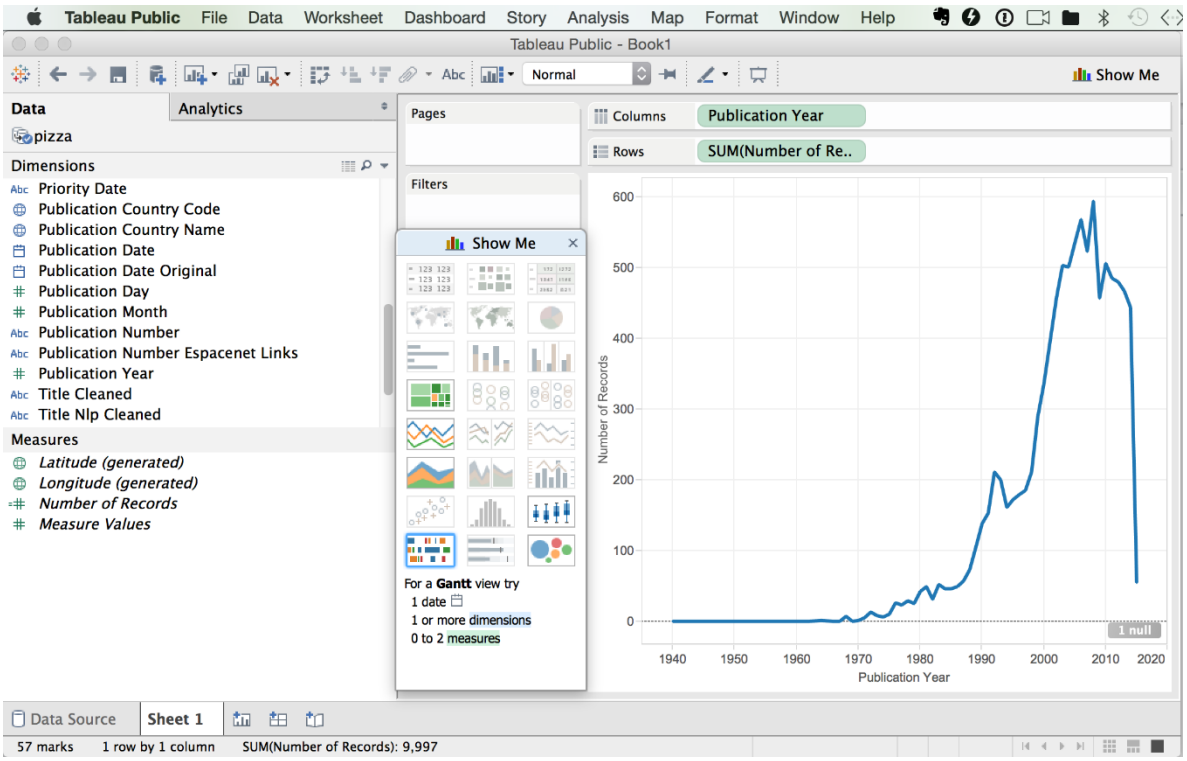
Por el momento, nos centraremos en crear hojas de trabajo con los datos y luego pasaremos a crear Cuadros de mandos y luego Historias en torno a nuestros datos de pizza.

9.4 Tendencias de publicación

Una de las primeras cosas que normalmente queremos hacer con los datos de patentes es mapear las tendencias, ya sea en las primeras presentaciones, publicaciones o miembros de la familia. En el caso de nuestras patentes de pizza de Patentscope tenemos un solo miembro de un expediente de archivos vinculados a una aplicación en particular. Estos datos están bien para las necesidades de demostración y podemos mapear fácilmente las tendencias para estos datos.

Para hacerlo, simplemente arrastramos el año de publicación en las dimensiones al campo de columnas y el número de registros del campo de medidas. Tenga en cuenta que Tableau cuenta automáticamente el número de filas en un conjunto para crear este campo. Si trabaja con datos en los que los recuentos precisos son importantes, es importante asegurarse de que los datos se hayan deduplicado en el campo correspondiente antes de comenzar. Si bien no se aplica en este caso, otra sugerencia importante es tener siempre una forma de verificar los recuentos clave en Tableau, como el uso de tablas dinámicas rápidas en Excel u Open Office. No tenemos que preocuparnos por esto ahora, pero si bien Tableau es un software inteligente, sigue siendo un software: no siempre realizará los cálculos como usted los espera. Por esa razón, una verificación cruzada de los recuentos es una parte sensible, si no vital, de un flujo de trabajo de Tableau.

Tableau adivinará lo que buscamos y dibujará una gráfica.



Como podemos ver, ahora tenemos un gráfico que se precipita desde un acantilado a medida que nos acercamos al presente y contiene uno nulo. Los valores nulos suelen ser filas o columnas que contienen celdas en blanco. Si solo hay 1 valor nulo, es probable que los datos se puedan dejar como están (en este caso era una fila en blanco en la parte inferior del conjunto de datos introducido durante la limpieza en R). Sin embargo, vale la pena inspeccionar los valores nulos haciendo clic derecho en el archivo Datay seleccionando View data. Si hay un gran número de NAvalores nulos, es posible que deba retroceder e inspeccionar los datos y asegurarse de que las celdas en blanco estén llenas de valores. Volvamos a nuestra gráfica.

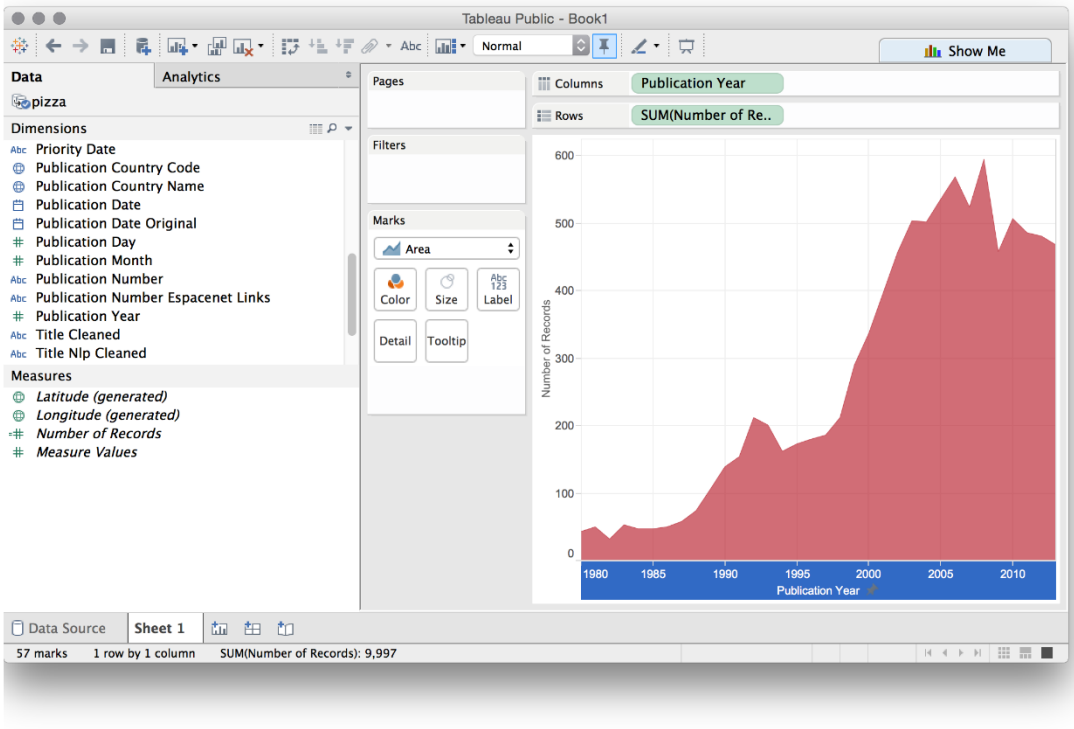
Lo que vemos aquí es lo data Cliff que es común con los datos de patentes. Es decir, el precipicio no representa un declive radical en el uso del término pizza, representa un declive radical en la disponibilidad de datos de patentes cuanto más nos acercamos al presente. La razón de esto es que generalmente, como regla general, se demora unos 24 meses para que se publique una solicitud y puede llevar más tiempo para que las bases de datos de patentes se pongan al día. Como tal, nuestro data Cliff refleja una falta de datos disponibles en los últimos años, no una falta de actividad. Por lo general, necesitamos retroceder unos 2 o 3 años para obtener una impresión de la tendencia.

Antes de continuar y ajustar el eje, cambiaremos el gráfico a algo más atractivo. Para ello seleccionaremos el gráfico relleno en el panel flotante. Detrás de ese

panel hay un pequeño botón de color que nos permitirá seleccionar un color que nos guste. La razón por la que hacemos esto antes de ajustar el eje es que cuando cambiamos el tipo de gráfico, Tableau revertirá cualquier cambio realizado en el eje.

A continuación, hacemos clic con el botón derecho en el eje x (inferior) y ajustamos el marco de tiempo a algo más razonable, como 1980 a 2013, seleccionando la fixed opción. Como una regla general muy aproximada, retroceder dos o tres años a partir del presente eliminará el precipicio de datos de la falta de información de patentes publicada. Tenga en cuenta que si contáramos las primeras solicitudes (familias de patentes), la disminución sería más temprana y mucho más pronunciada. El equipo de estadísticas de patentes de la OCDE ha investigado detalladamente estos efectos de retraso y las formas de tratarlos , ver el trabajo en particular sobre los datos de patentes de difusión inmediata .

Ahora tenemos una buena gráfica con un eje sensible. Tenga en cuenta que si estuviésemos graficando múltiples tendencias en el mismo gráfico (familia y miembros de la familia), podríamos preferir un gráfico de líneas sencillo en aras de la claridad.

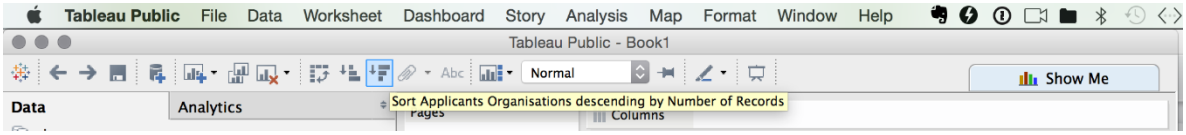


Le daremos un nombre a esto Trendsy agregaremos una nueva hoja de trabajo haciendo clic en el ícono junto a nuestra hoja existente.

La siguiente información que nos gustaría es quiénes son los solicitantes más activos. Esto también comenzará a exponer temas sobre los diferentes actores que utilizan el término pizza en el sistema de patentes y nos alentará a pensar en formas de profundizar en los datos para obtener información más precisa sobre las tecnologías que nos pueden interesar, como en este caso. , cajas de pizza y cajas de pizza musicales en particular.

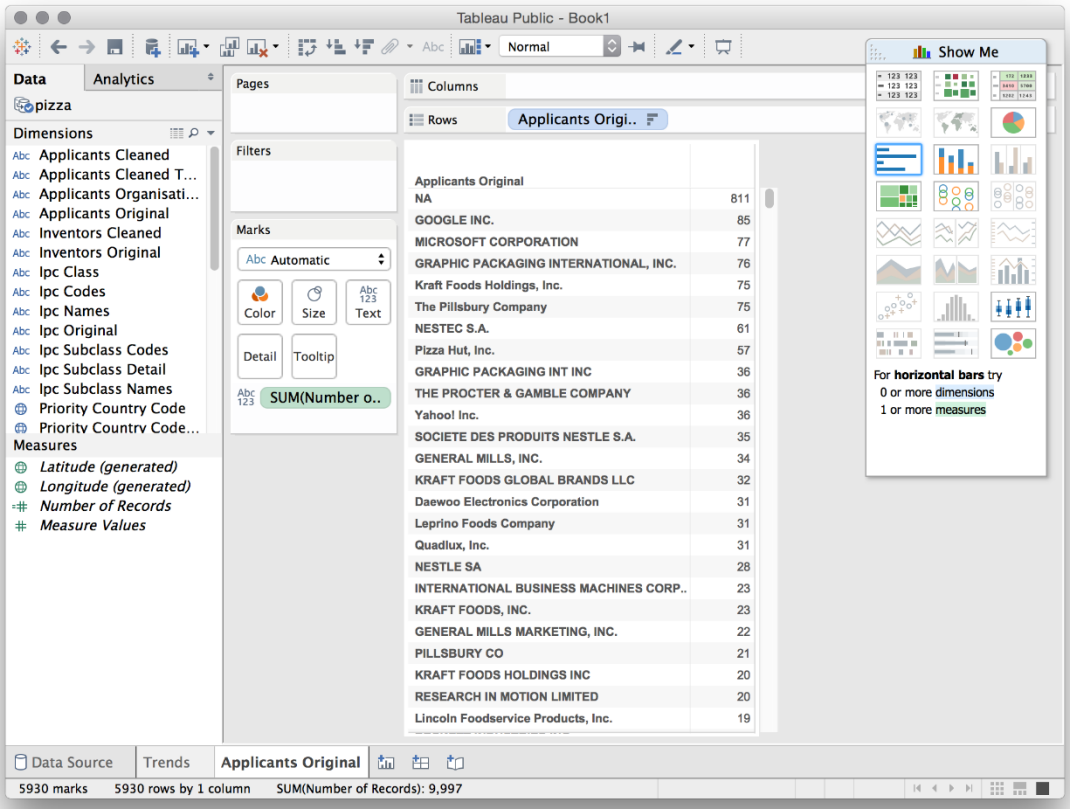
Es en este punto que el trabajo que hicimos en un artículo anterior sobre cómo separar los nombres de los solicitantes individuales en sus propias filas y cómo limpiarlos con Open Refine, se vuelve importante. En este conjunto de datos, hemos llevado esto un paso más allá utilizando VantagePoint para separar a las personas de las organizaciones. Esta información se encuentra en el Applicants Organisations campo en el conjunto de datos. Simplemente coloquemos eso en la hoja de cálculo como una fila y luego agregamos el número de registros como una columna (sugerencia, simplemente suéltelo en la hoja).

A primera vista todo parece bastante bien. Pero ahora tenemos que clasificar a nuestros aspirantes. Para hacer eso, seleccionamos el pequeño icono en la barra de menú con una barra apilada apuntando hacia abajo.



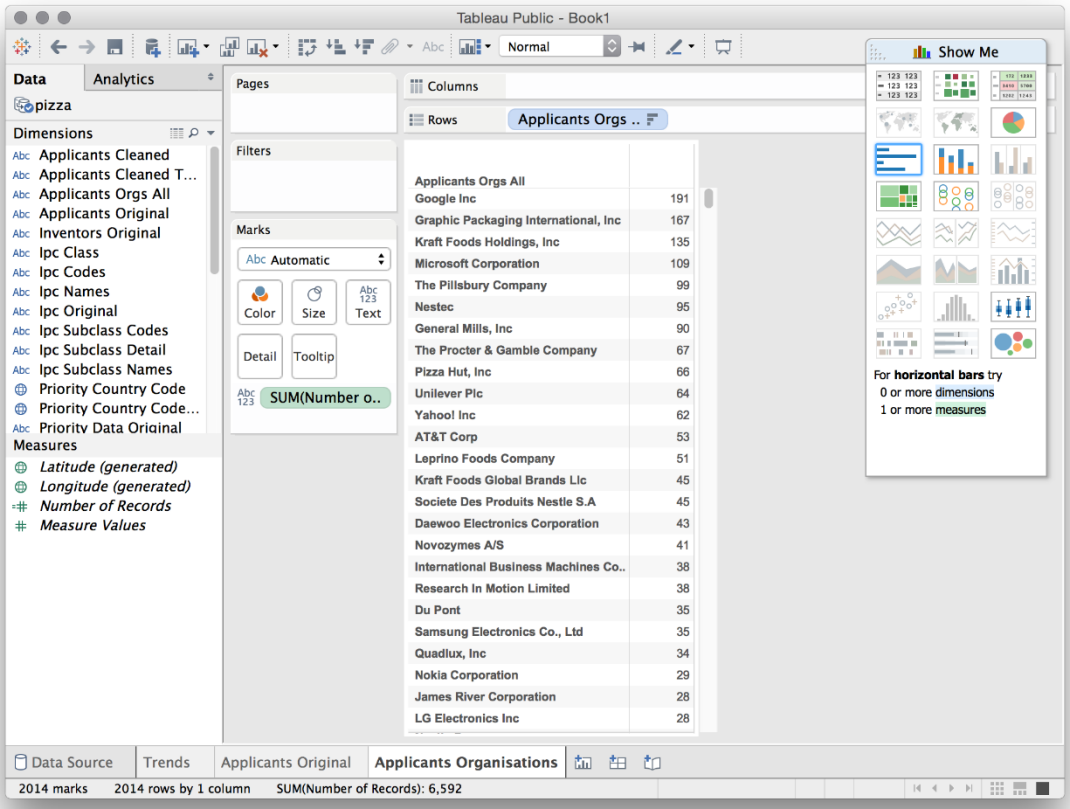
Ahora vemos, como veríamos en el archivo sin formato de Excel, que hay un número significativo de entradas en blanco para los solicitantes en los datos subyacentes, seguidos por 85 registros para Google y 77 para Microsoft. Este también es un muy buen indicador de que puede haber múltiples usos de la palabra pizza en el sistema de patentes, a menos que estas compañías de software hayan comenzado a vender pizzas en línea.

En realidad, esta es ***una vista parcial de la actividad*** por parte de los solicitantes porque en otra parte de los datos los nombres se concatenan juntos. Esto suele ser más obvio que en el conjunto de datos actual a través de la presencia de varios nombres separados por ;(para ver este desplazamiento hacia abajo hasta la primera entrada de Unilever).



Para entender por qué esta es una vista parcial, ahora importaremos el applicants.csvarchivo. La forma correcta de hacerlo es seleccionar el menú llamado Dataa continuación New Data Sourcey el archivo applicants.csv.

A continuación, arrastre Applicants Orgs Allhacia las filas. Tenga en cuenta que Tableau está interpretando estos títulos para nosotros (el original es applicants_orgs_all). Luego arrastre Number of Recordsdesde las dimensiones a la hoja o a la entrada de columnas. Ahora elija el icono de la barra apilada como se indica arriba para clasificar a los solicitantes por el número de registros. Ahora veremos lo siguiente.

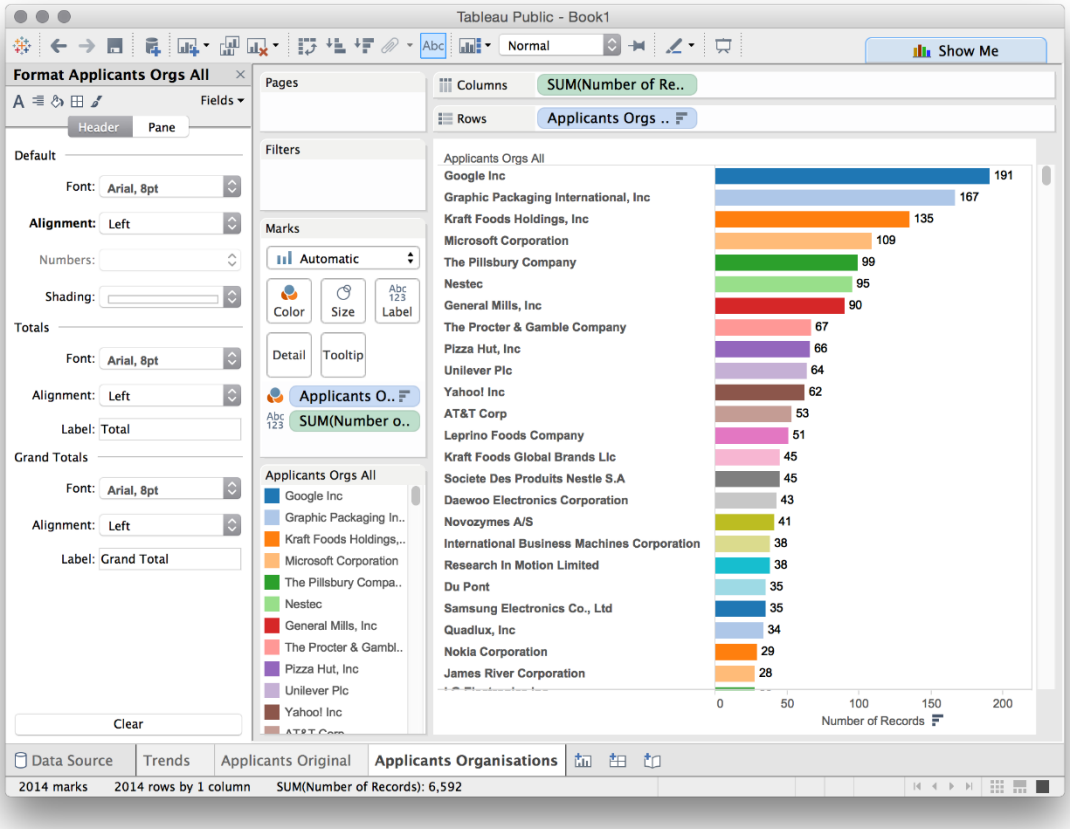


Tenga en cuenta la diferencia entre el campo de los solicitantes originales (donde Google obtuvo un total de 85 registros) y nuestro campo separado y limpio donde Google ahora obtiene 191 registros. En resumen, antes de los ejercicios de separación y limpieza solo vimos el 44% de la actividad en nuestro conjunto de datos de Google con el término pizza. Esto todavía no significa que hayan entrado en el negocio de la pizza en línea ... Lo que sí nos dice es que el análisis de patentes que no separa ni divide los datos concatenados y las variantes de nombres de limpieza falta más del 60% de la historia cuando se ve en términos de la actividad del solicitante. Como queda claro, los beneficios de separar o dividir y limpiar los datos son enormes incluso cuando, como en este caso, los datos originales parecían ser bastante "limpios". Esa apariencia era engañosa.

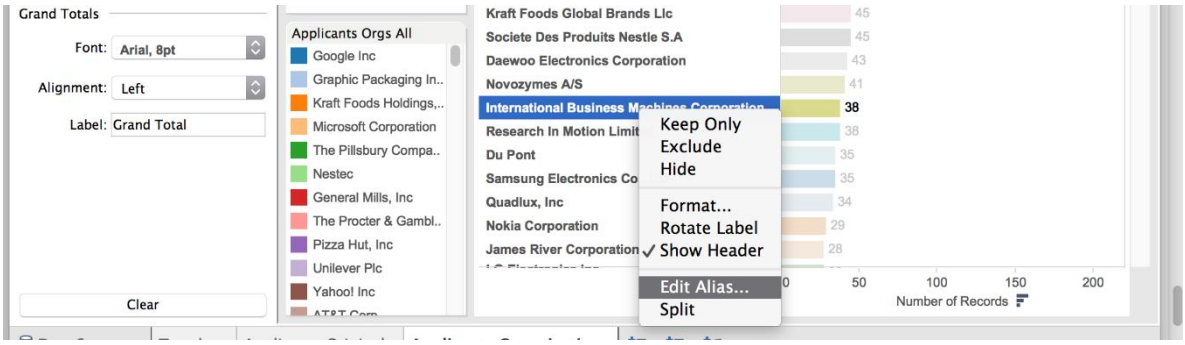
Ahora que tenemos una visión más clara de lo que está sucediendo con nuestros solicitantes, podemos hacer que esto sea más atractivo. Para hacer eso primero selecciona la barra azul en el panel flotante. La hoja de trabajo ahora se presentará como barras clasificadas. A continuación, arrastre el número de registros desde Medidas al Labelbotón al lado de Color. Eso se ve bastante bien. Si quisiéramos ir un paso más allá, ahora podríamos pasar al panel de dimensiones y arrastrar Applicants Orgs All, al Colorbotón. Las barras ahora cambiarán a colores

diferentes para cada solicitante. Si esto es demasiado brillante, simplemente tome el Applicants Orgs Allcuadro verde debajo del menú de botones y muévalo hacia las dimensiones para eliminarlo. Finalmente, si queremos ajustar la alineación derecha del texto a la izquierda, primero haga clic derecho en el nombre de una empresa, seleccioneFormatluego la alineación y la izquierda. Mientras que el valor predeterminado es la alineación a la derecha, en la práctica la alineación a la izquierda crea más etiquetas legibles. Para cambiar el valor predeterminado, haga esto con la primera hoja de trabajo que cree antes de crear cualquier otra.

Ahora tenemos una tabla de datos de solicitantes que se ve, dependiendo de su sensibilidad estética, como esta.



En esta etapa podríamos querer tomar un par de acciones. Para hacer que las etiquetas sean más visibles, arrastre la línea entre los nombres y las columnas a la derecha. Esto abrirá un poco de espacio. A continuación, piense en editar nombres largos en algo corto. Por ejemplo, International Business Machines Corporation, que tampoco es famosa por las pizzas, es demasiado larga. Haga clic derecho en el nombre y seleccione Edit aliascomo en la imagen de abajo.



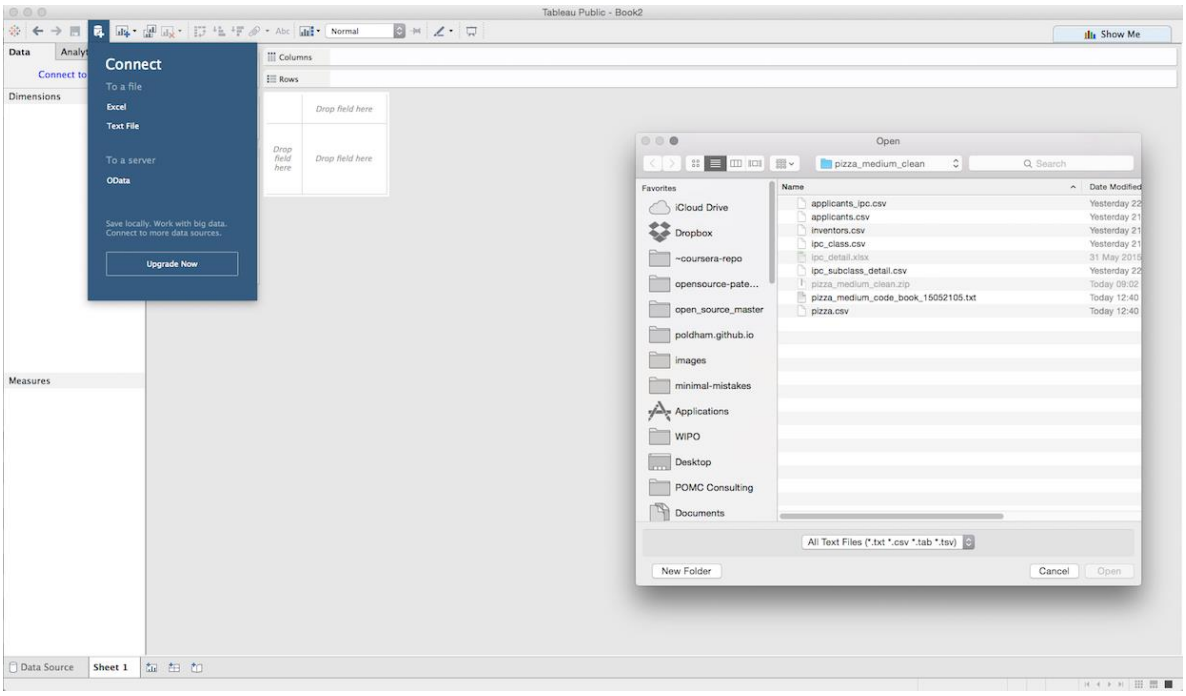
Ahora edita el nombre a IBM. Como sugerencia, cuando descubra que ha omitido un nombre duplicado en la limpieza (recuerde que nos centramos en lo suficientemente bueno en lugar de ser perfecto en la limpieza de datos) también es posible resaltar dos filas, hacer clic con el botón derecho, buscar un icono de clip de archivo y agrupar dos entradas en un nuevo nombre. Sin embargo, el grupo nombrado resultante debe usarse en todos los análisis posteriores. También es importante darse cuenta de que la limpieza de datos no es una fortaleza de Tableau, Tableau se trata del análisis de datos y la exploración a través de la visualización. Para la limpieza de datos use una herramienta como Open Refine.

9.5 Agregando nuevas fuentes de datos

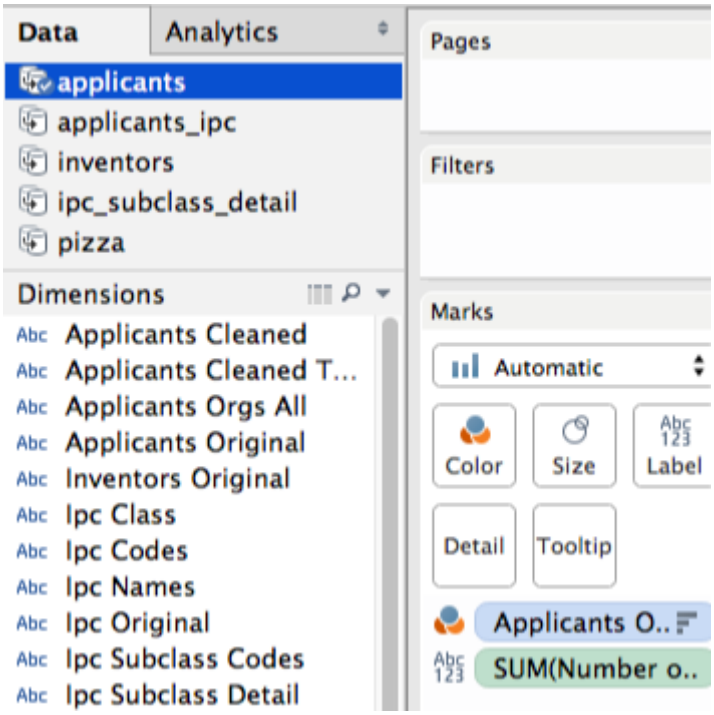
Seguiremos el mismo procedimiento que utilizamos para que los solicitantes agreguen los archivos restantes como fuentes de datos. Agregaremos los siguientes cuatro archivos (tal como aparecen en la carpeta en orden alfabético).

1. solicitantes_ipc.csv
2. inventors.csv
3. ipc_class.csv
4. ipc_subclass.detail.csv

Para agregar las fuentes de datos, haga clic en el Datamenú y New Data Sourceo (más rápido) el cilindro con un signo más. Luego seleccione Text file, agregue cada archivo y permita que se cargue.



Si todo va bien, el Data panel ahora contendrá los siguientes archivos.

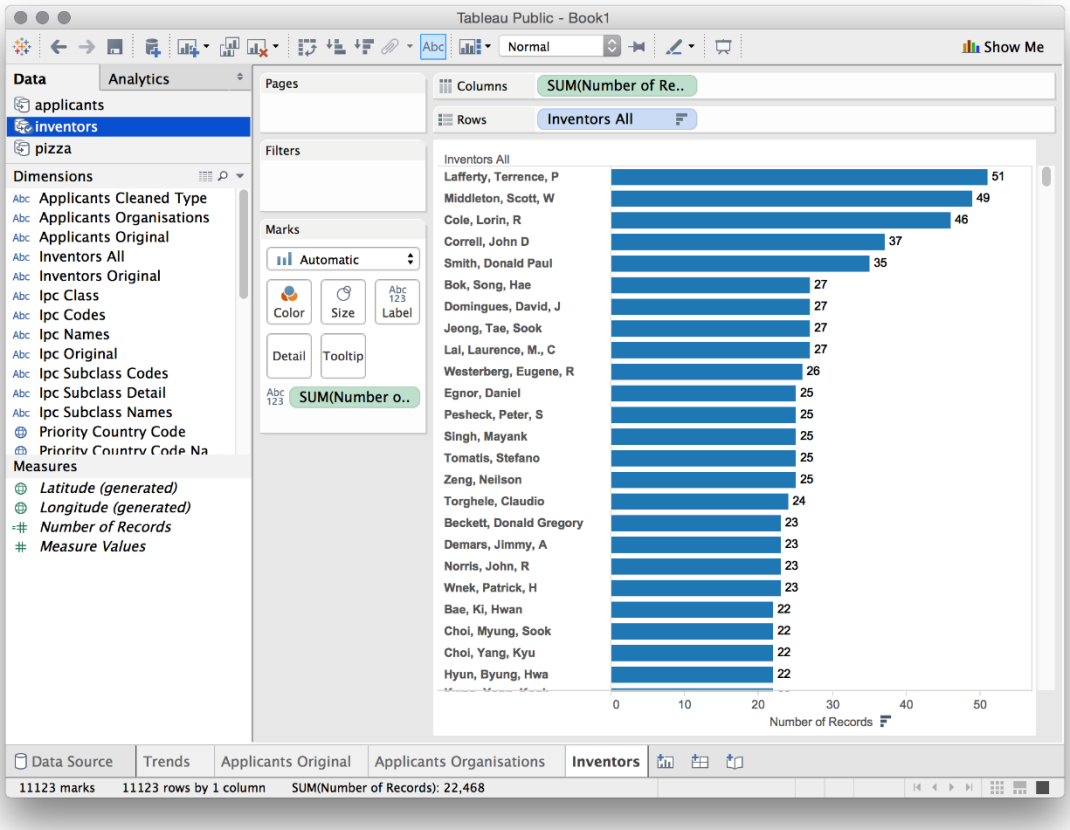


Tenga en cuenta que los applicants datos muestran una marca azul. Esto se debe a que fue la última fuente de datos que usamos y, por lo tanto, está activa. Los campos que vemos en Dimensiones pertenecen a esa fuente de datos. A continuación, haga clic en el menú inferior para crear una nueva hoja de cálculo y

luego haga clic inventors en el Data campo. Los nombres de los campos ahora cambiarán ligeramente. Es importante vigilar la fuente de datos que está utilizando porque es muy fácil colocar un campo de una fuente de datos en otra. En algunos casos esto es algo bueno. Pero, si recibe un mensaje de advertencia, intentará colocar una fuente de datos en otra fuente de datos donde no haya un campo coincidente. Volveremos a esto en los datos blending.

A continuación, siga el mismo procedimiento para clasificar a los solicitantes con los inventores que utilizan el Inventors All. Para cualquier persona interesada en ver los dramáticos impactos de los campos concatenados, intente colocar el Inventors Original campo en la hoja de trabajo.

Usando Inventors All ahora debería ver la siguiente lista clasificada de inventores.



Ahora repita este ejercicio para las fuentes de datos restantes creando primero una hoja y luego seleccionando la fuente de datos. A medida que avance a través de este, seleccione las siguientes dimensiones para agregar a la hoja y luego suelte el número de

1. solicitantes_ipc. Caer Ipc Subclass Detailsobre la hoja. Luego suelta el número de registros en la hoja donde el campo dice Abc. Tenga en cuenta que 6aparecerá un número en la primera fila. Este es un artefacto del proceso de separación. Seleccione esa celda, haga clic derecho y luego elija Exclude.

No clasifique estos datos, sino que arrastre el campo Applicants Orgs Alla la hoja para que sea la primera fila (sugerencia, es más fácil hacerlo arrastrando el campo a la barra de la fila antes del campo IPC). Ahora verá una lista de nombres de compañías seguidas de una lista de IPC. Enhorabuena, ahora tenemos una idea de quién está patentando en un área particular de la tecnología usando la palabra pizza a nivel de solicitantes individuales.

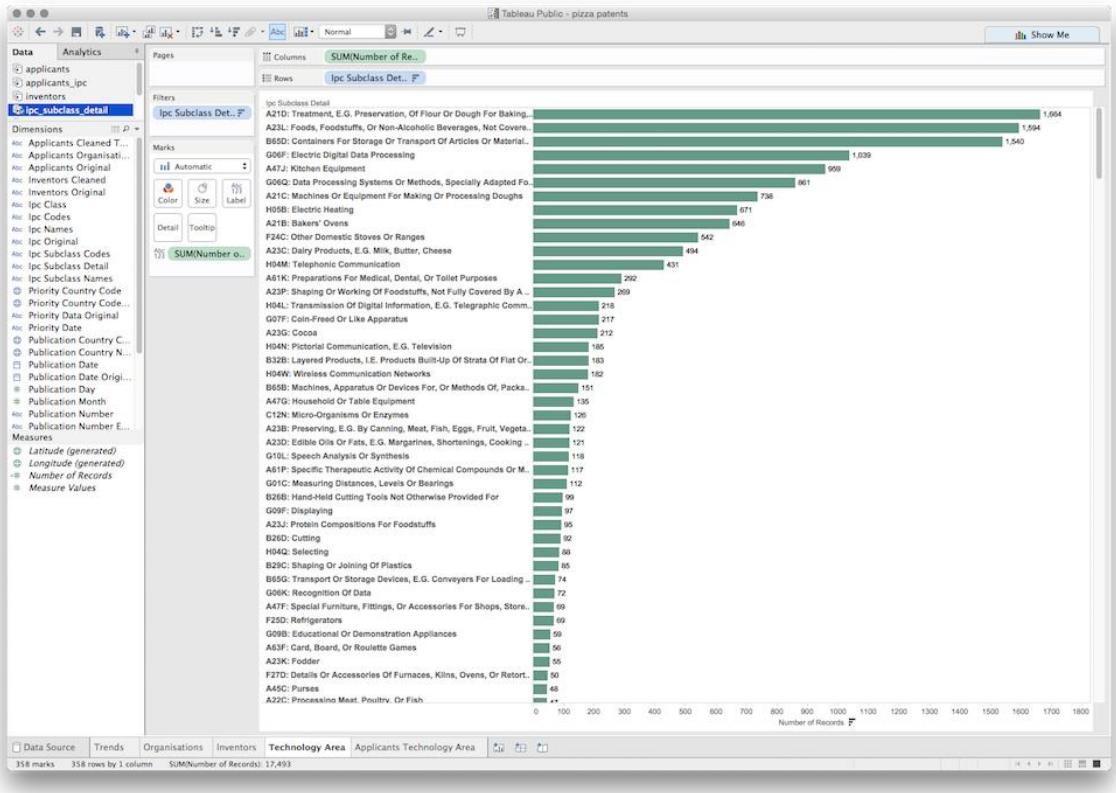
Añadir una nueva hoja. Luego haga clic en ipc_subclass_detail. Tenga en cuenta que si hace clic en la fuente de datos primero, el panel de dimensiones se pondrá de color naranja. No se asuste. La razón es que Tableau cree que está intentando combinar datos de la fuente ipc_subclass_detail con Applicants_ipc. Si haces esto, simplemente haz clic en ipc_subclass-detail otra vez.

2. ipc_subclass-detail. Coloca la Ipc Subclass Detail dimensión en la hoja. Luego suelte el número de registros en la hoja. Luego haga clic en la primera celda que contiene 6como artefacto y excluya. Repita para 7. Luego, seleccione el gráfico de barras en el Show Mepanel flotante, luego arrastre Number of Records hacia el Label botón. Ahora clasifique la columna usando el botón descendente en el menú superior como antes.

En este punto, si no hubiéramos recortado el espacio en blanco inicial, la lista clasificada mostraría sangrías y habría duplicados del mismo código IPC. Por esa razón, es importante recortar los espacios en blanco iniciales antes de intentar visualizar los datos (y esto se aplica a todos nuestros campos separados).

9.6 Creación de un cuadro de mando general

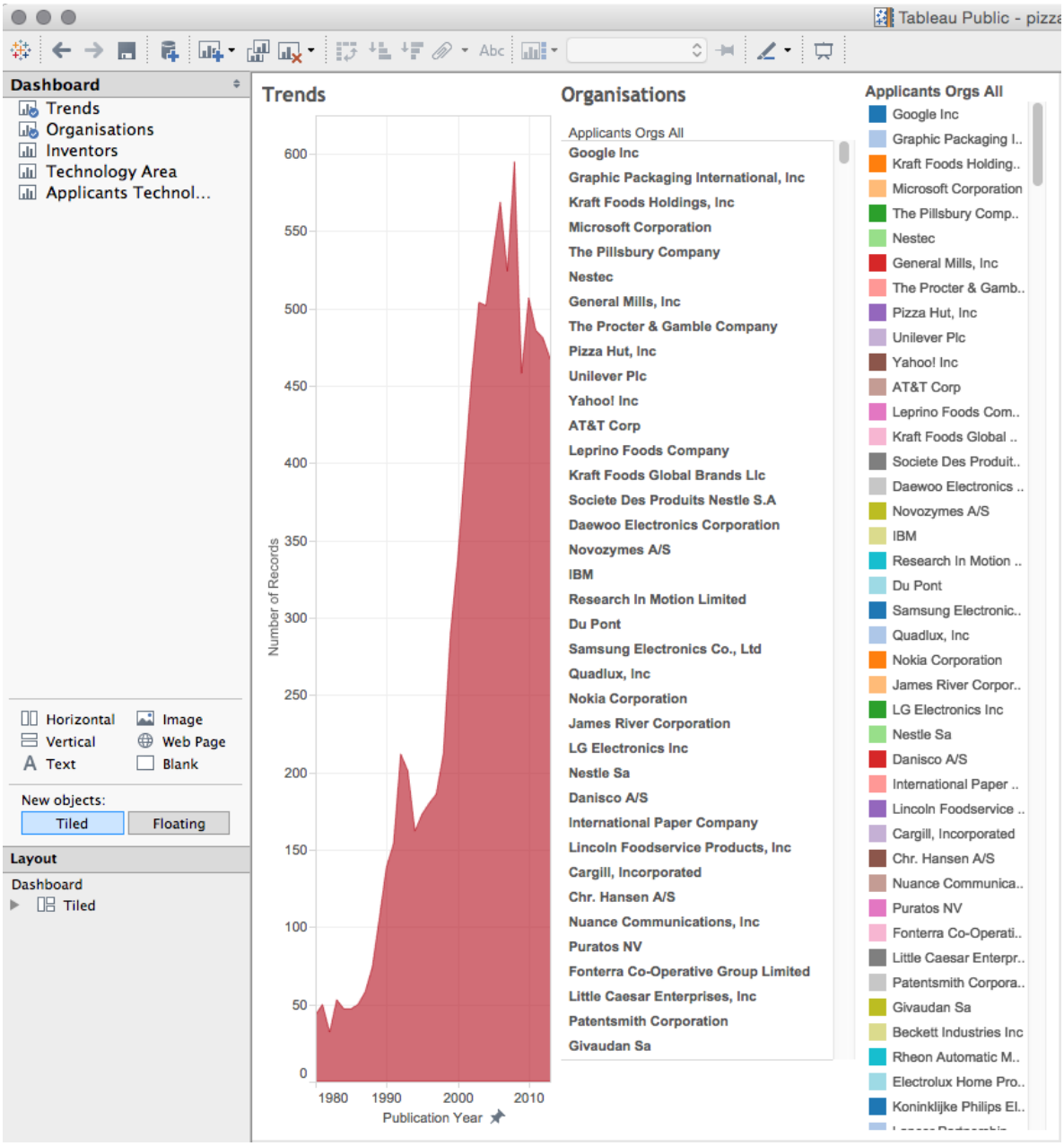
Ahora debería tener cinco hojas de trabajo, cada una de las cuales muestra aspectos de nuestro pizza conjunto principal. Hemos nombrado las hojas de la siguiente manera y le sugerimos que desee hacer lo mismo. Tenga en cuenta que cuando haya más de una hoja que contenga información similar pero distinta, será útil darles nombres distintos (por ejemplo, Subclase de IPC y Subclases de IPC de solicitantes). Incluso podríamos comenzar a usar etiquetas menos técnicas al llamar al IPC algo más claro, como el Área de Tecnología, para ayudar a la comunicación con especialistas que no son de IP.



Vamos a obtener una visión general rápida de los datos hasta el momento. Junto al botón Agregar hoja de trabajo en la barra de hojas de trabajo hay un segundo icono para crear un panel. Haga clic en eso y ahora veremos una hoja llamada Dashboard 1.

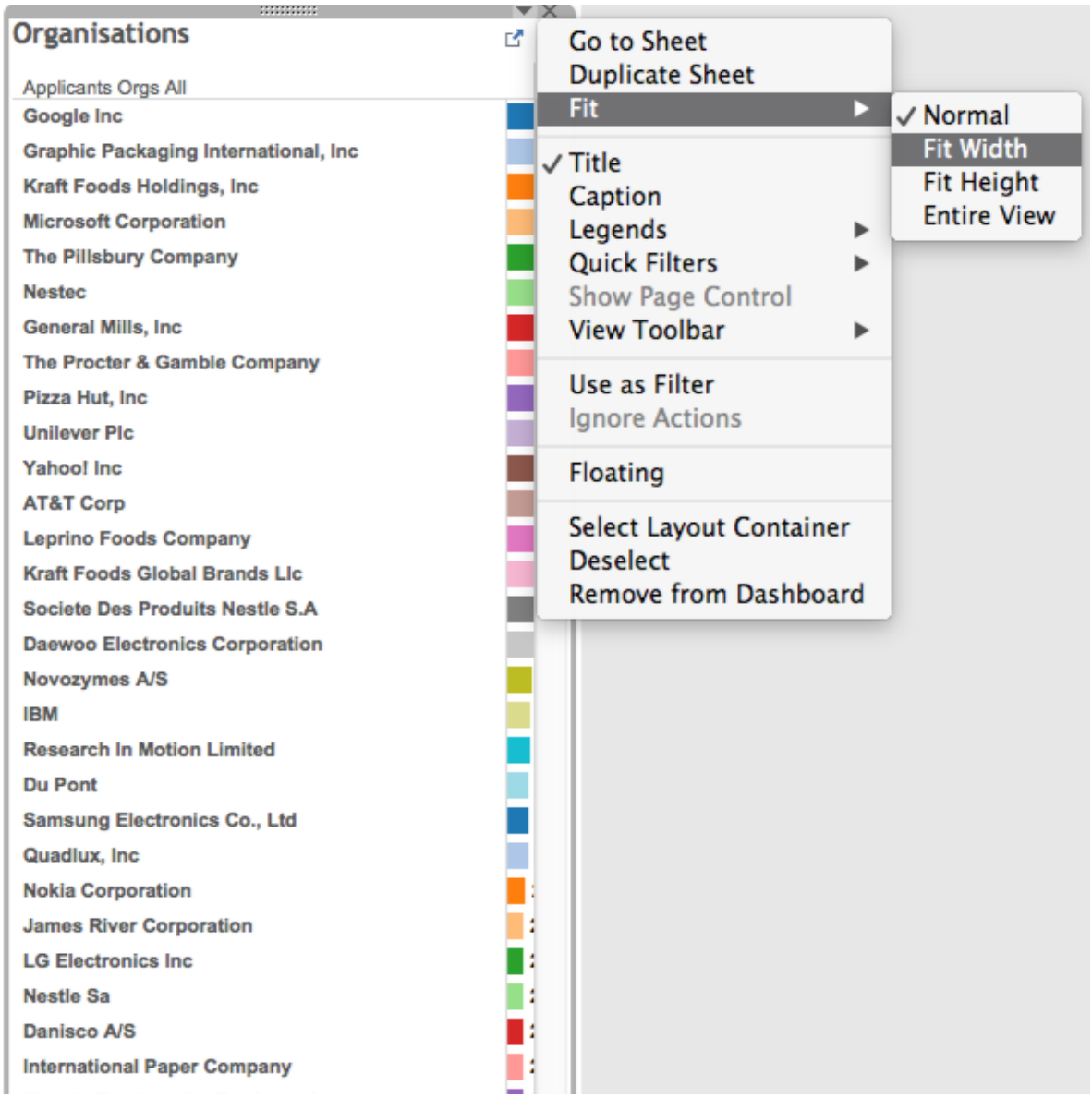
Los paneles de control son quizás la característica más conocida de Tableau y, con razón, son muy populares. Podemos llenar nuestro panel de control arrastrando las hojas de trabajo desde el Dashboard menú lateral. El orden en el que hacemos esto puede hacer la vida más fácil o más difícil de ajustar más adelante. Hagámoslo en los siguientes pasos.

1. Arrastre Trends hacia el tablero de mandos y ahora llenará la vista.
2. Arrastre Organisations en el tablero de mandos.



Eso es bastante desordenado, pero no todo está perdido. Simplemente haga clic en la esquina superior derecha del panel de organizaciones a la derecha para eliminarlo (en la hoja de trabajo original, haga clic en él y seleccione Hide). Ahora tenemos una Organisations columna que todavía parece crujida.

Ahora seleccione la parte superior del cuadro de organizaciones y aparecerá un pequeño triángulo invertido. Haz clic en eso y luego elige Fit > Fit Width.



Las barras ahora pueden desaparecer. Haga clic en el cuadro en la línea donde comienzan las barras y arrástrelas nuevamente a la vista. En este punto, los nombres largos pueden comenzar a ocultarse. Si lo desea, haga clic con el botón derecho en un nombre largo, por ejemplo Graphic Packaging International, elíjalo Edit aliasy edítelo a un valor razonable, como Graphic Packaging Int.

Ahora tenemos dos paneles en el salpicadero. Vamos a añadir dos más. Primero arrastre las áreas de tecnología debajo de la línea donde terminan Tendencias y Organizaciones. Aparecerán cuadros sombreados en gris que muestran la ubicación, en todo el ancho está bien. Esto puede tomar algún tiempo para hacerlo bien, cuando toda la zona inferior está resaltada, suelte el mouse. Si va a algún lugar extraño, seleccione el cuadro y, en la parte superior derecha, presione xpara

eliminarlo o intente moverlo (en nuestra experiencia, a menudo es más fácil eliminarlo e intentarlo de nuevo). No intente formatear este cuadro todavía. En su lugar, atrape a los inventores y arrástrelos al espacio antes de las áreas de tecnología a continuación.

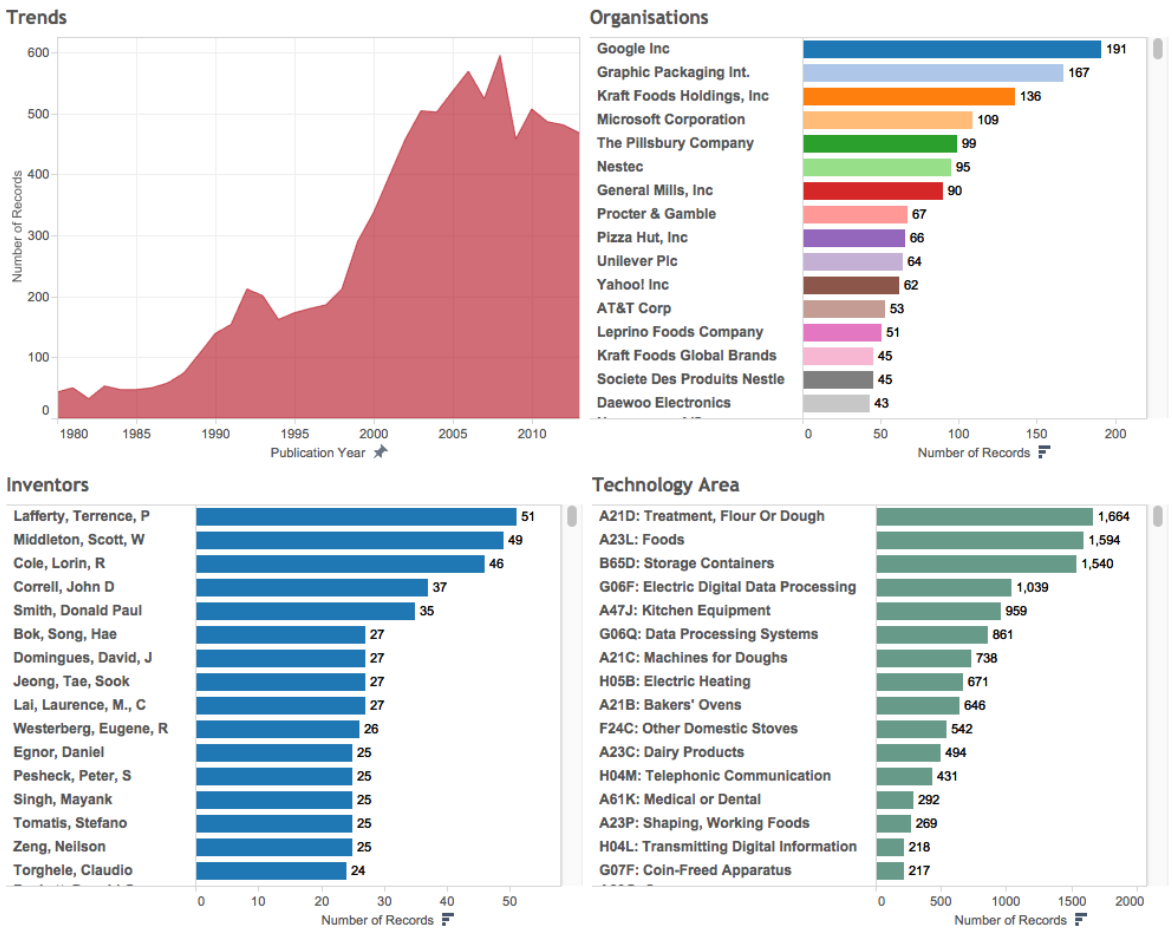
Ahora tenemos cuatro paneles en el tablero de mandos, pero necesitan un poco de ordenación. Primero, en los dos cuadros que acabamos de editar repita el Fit Width ejercicio y luego arrastre la línea de las barras hasta que estén a la vista y sean satisfactorias. A continuación, tenemos nombres como los Applicants Orgs All que son nuestros nombres de referencia internos. Haga clic en ellos en cada uno de los tres paneles uno a la vez y seleccione Hide Field Labels for Rows.

Hmm ... nuestro panel de áreas tecnológicas está resultando problemático porque incluso la versión editada del IPC es bastante larga.

Antes de realizar cualquier edición, primero experimente con el Size menú en la parte inferior derecha. El tamaño de panel predeterminado en Tableau Public es bastante pequeño. Cambie la configuración hasta que tenga algo que se vea más limpio incluso si todavía hay algunas superposiciones. Las opciones como Desktop, Laptopy en Large blog general son tamaños decentes, pero en parte la decisión depende de dónde cree que se mostrará.

Para corregir las etiquetas largas de áreas tecnológicas, volvemos a la hoja original (sugerencia: si mueve el mouse hacia la parte superior derecha en el panel Go to Sheet aparecerá una flecha, es muy útil para libros grandes). Dentro de la hoja original, intente arrastrar la línea que separa el texto y las barras para que las barras ahora cubran parte del texto más largo. A continuación, vuelva al panel de control. Si no está satisfecho con el resultado, haga clic con el botón derecho en el panel en el panel y luego elija Edit alias. Esto es útil para simplemente hacer que las etiquetas en la vista sean más visibles (no cambia los datos originales).

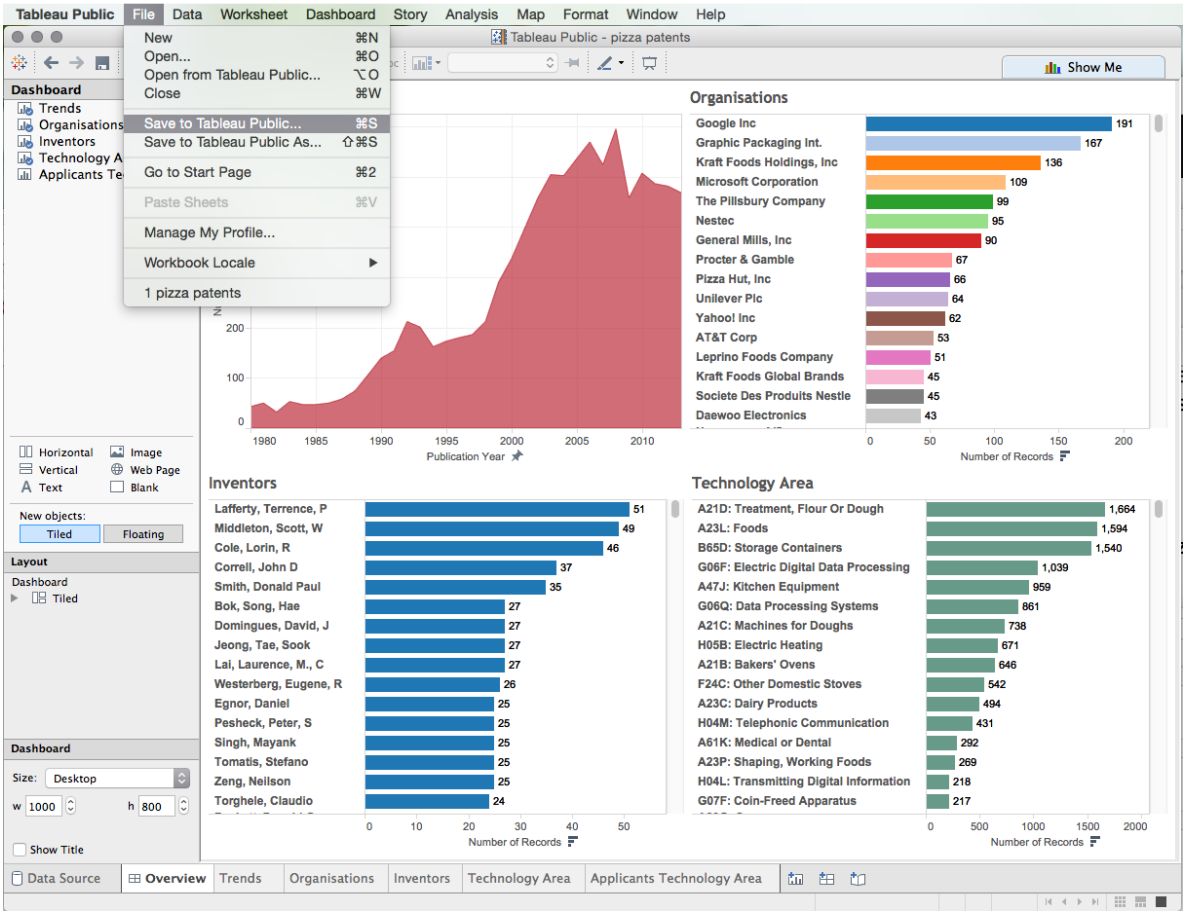
Si todo va bien, ahora tendrá un panel de control que se verá más o menos así. Tenga en cuenta que, dependiendo de la configuración de la hoja de trabajo, puede querer que el tamaño de la fuente sea consistente (haga clic con el botón derecho y elija Formato, luego el tamaño de fuente). Tenga en cuenta también que si aumenta el tamaño de la fuente (el valor predeterminado es 8 puntos), es posible que deba editar algunas de las etiquetas nuevamente.



Ahora hemos hecho bastante trabajo y hemos producido un panel de información general. Es hora de guardar el libro de trabajo en el servidor antes de hacer cualquier otra cosa.

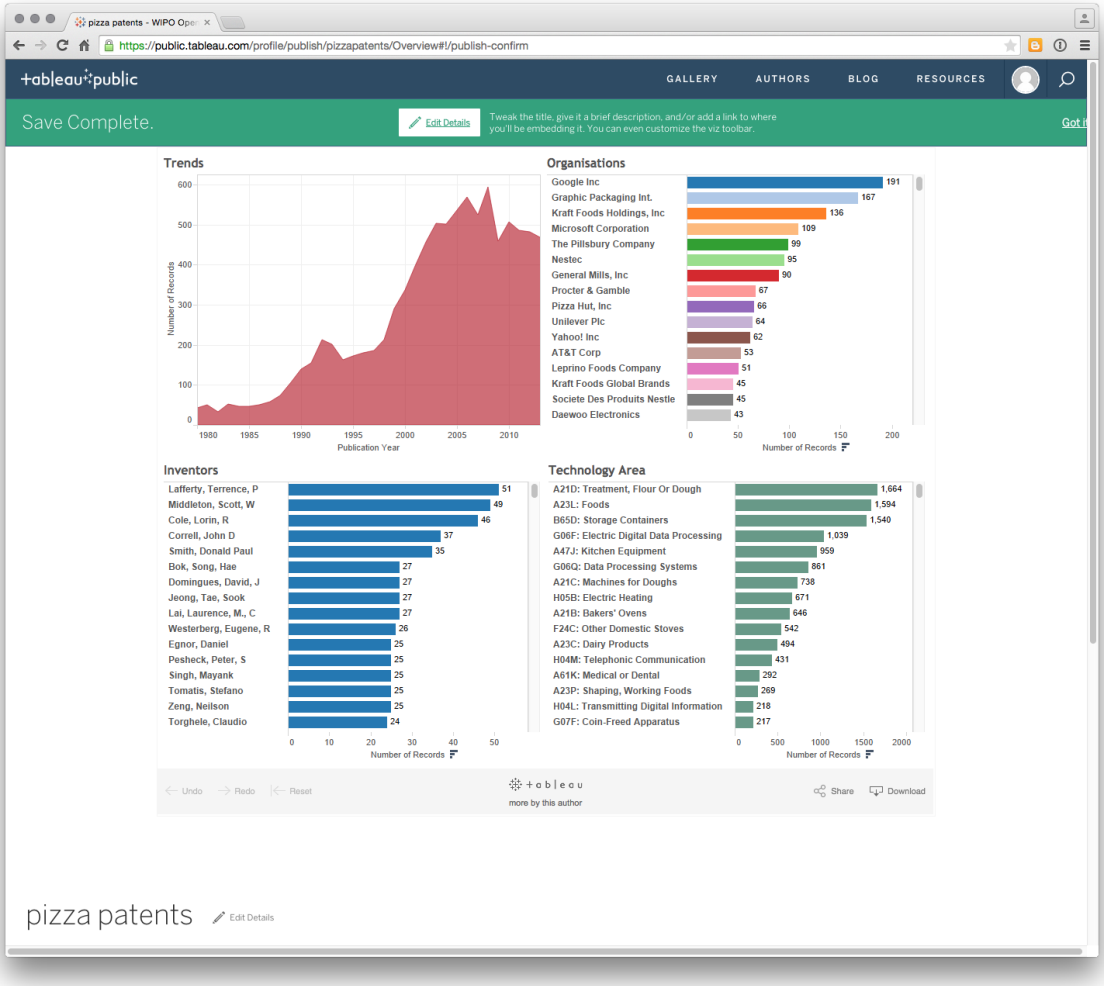
9.7 Configuración de guardado, visualización y privacidad

La única opción para guardar un libro de trabajo de Tableau Public es guardarlo en línea. Para guardar el archivo, vaya a File Guardar en Tableau Public. Si desea guardar el libro de trabajo como un archivo nuevo (después de guardar previamente), seleccione Guardar como.

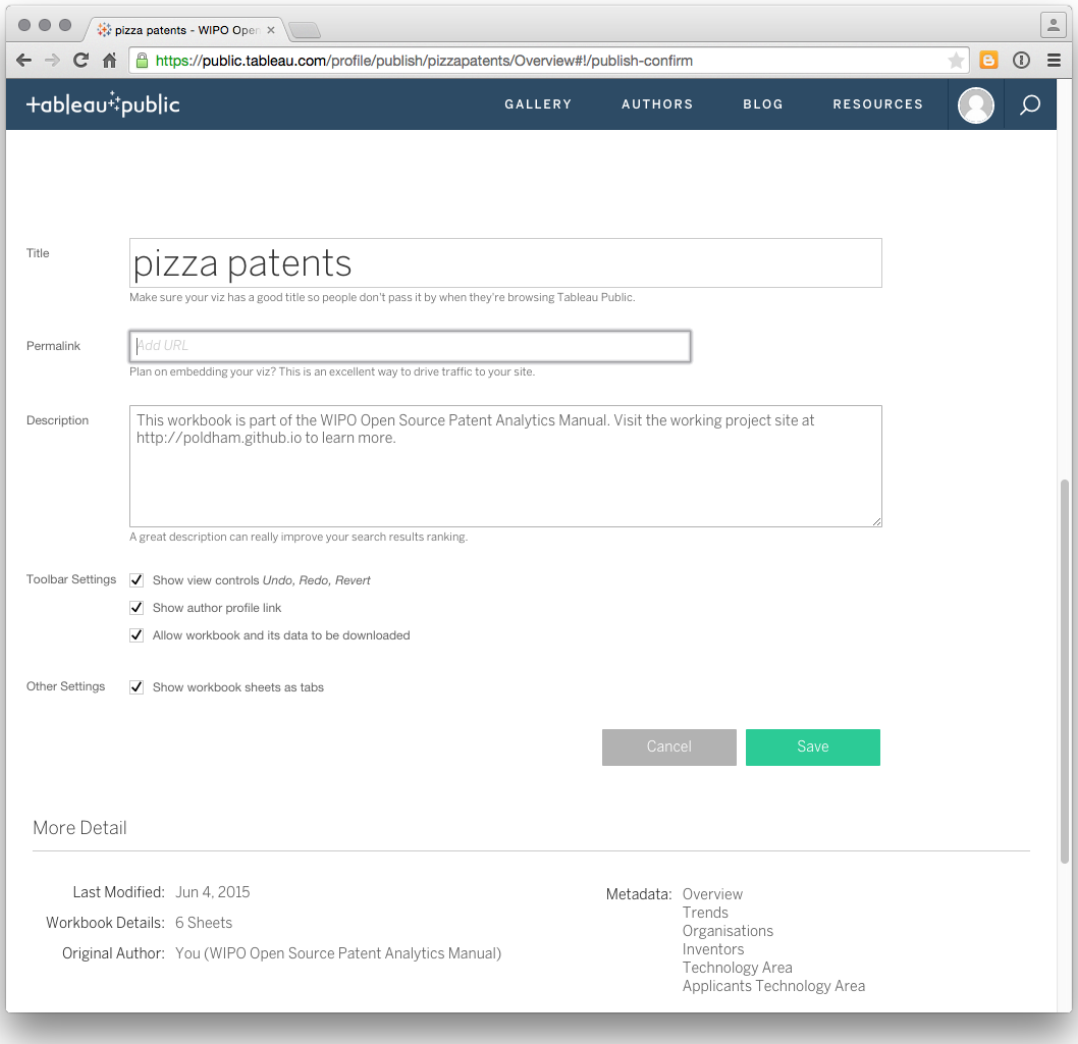


Luego se le pedirá que ingrese su nombre de usuario y contraseña (Tableau no recuerda la contraseña) y el archivo se cargará. Tableau luego comprimirá los datos. A partir de junio de 2015, es posible almacenar 10 GB de datos en general y tener hasta 10 millones de filas en un libro de trabajo (que generalmente es más que suficiente).

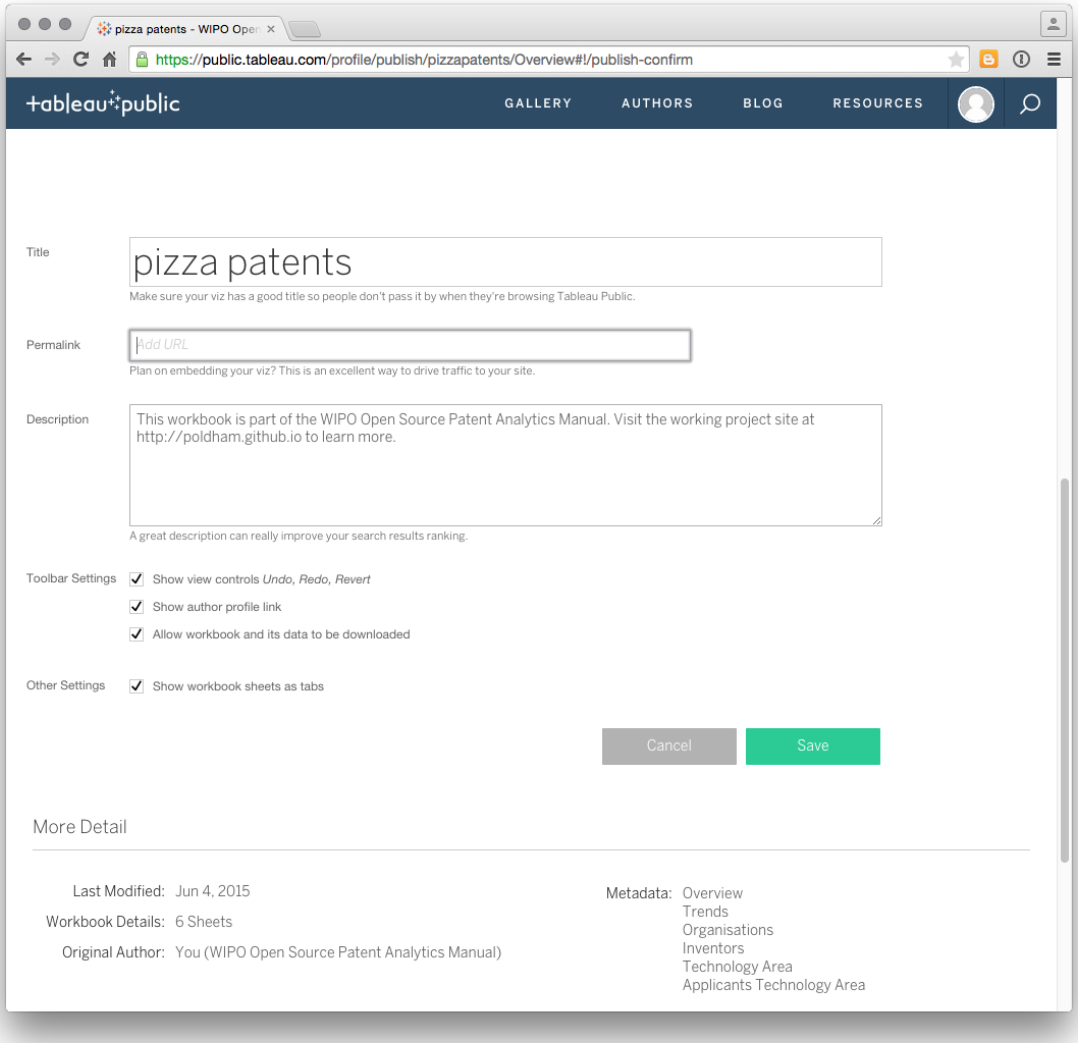
Tableau abrirá un navegador web en su página de perfil y se verá muy parecido a esto.



Después de leer el mensaje, haga clic Got it a la derecha. ¿Notan algo extraño? Sí, solo podemos ver el Panel y no ninguna de las otras hojas. Para cambiar este y otros detalles, haga clic en edit details cerca del título y se abrirán algunos menús de la siguiente manera.


pizza patents - WIPO Open
https://public.tableau.com/profile/publish/pizzapatents/Overview#!/publish-confirm
tableau public
GALLERY
AUTHORS
BLOG
RESOURCES
Title
pizza patents
Make sure your viz has a good title so people don't pass it by when they're browsing Tableau Public.
Permalink
Add URL
Plan on embedding your viz? This is an excellent way to drive traffic to your site.
Description
This workbook is part of the WIPO Open Source Patent Analytics Manual. Visit the working project site at http://poldham.github.io to learn more.
A great description can really improve your search results ranking.
Toolbar Settings
Show view controls Undo, Redo, Revert
Show author profile link
Allow workbook and its data to be downloaded
Other Settings
Show workbook sheets as tabs
Cancel
Save
More Detail
Last Modified: Jun 4, 2015
Workbook Details: 6 Sheets
Original Author: You (WIPO Open Source Patent Analytics Manual)
Metadata: Overview
Trends
Organisations
Inventors
Technology Area
Applicants Technology Area

Para asegurarse de que las hojas de trabajo estén visibles, seleccione la casilla de verificación marcada Show workbook sheets as tabs y luego Save.



Para acceder a este libro de demostración vaya aquí .

9.8 Privacidad y seguridad

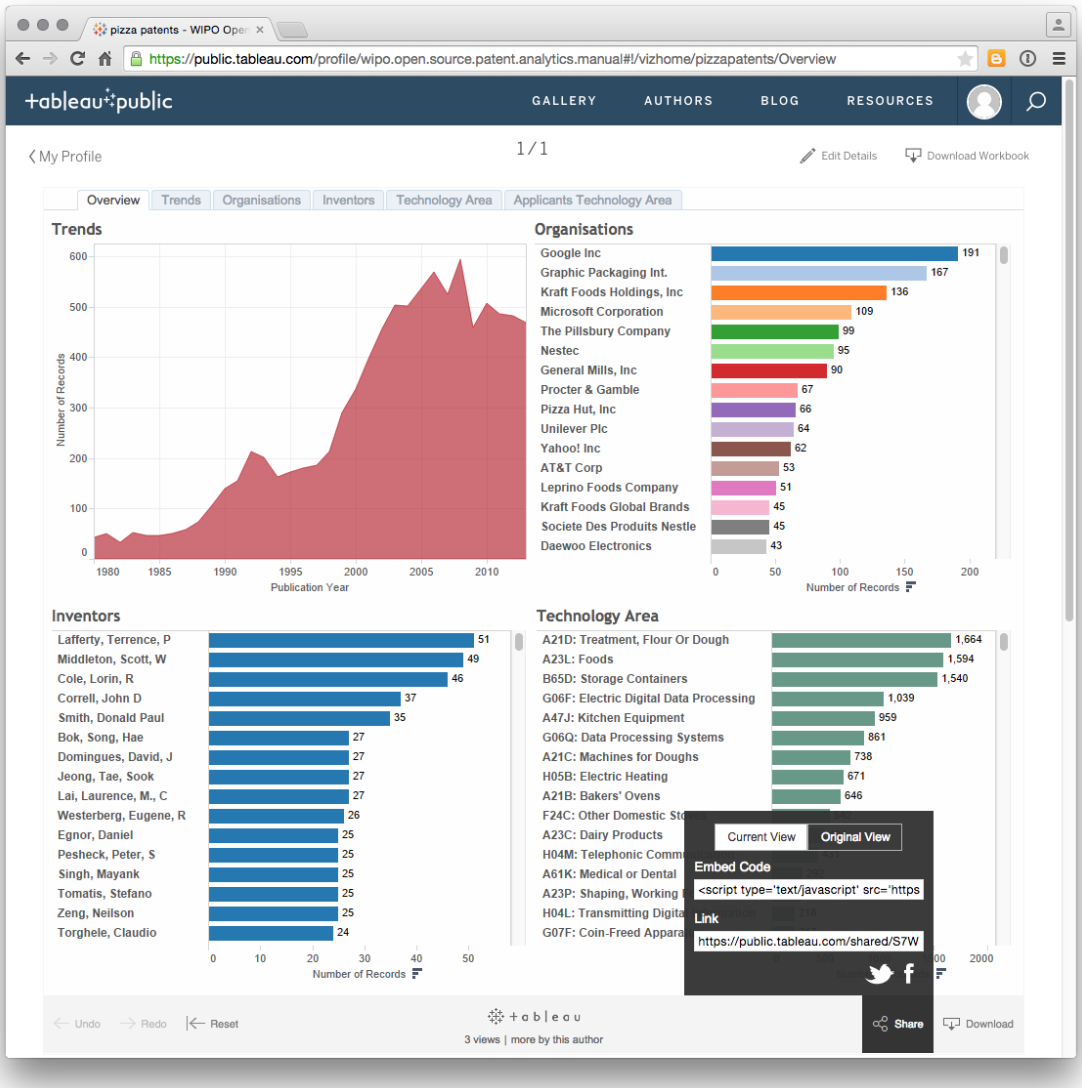
Como se destacó anteriormente, Tableau Public es, por definición, un lugar para compartir libros y visualizaciones públicamente. No es para datos sensibles. En el pasado, los usuarios, como los periodistas, confiaban en lo que podría llamarse "seguridad por oscuridad", pero la tendencia a almacenar datos en un perfil público de Tableau (la única opción) hace que sea una opción menos. Si esto es una preocupación, hay dos acciones que podrían considerarse que limitan la visibilidad de un libro de trabajo y su uso más amplio. Lógicamente, la respuesta a cualquier

inquietud sobre Tableau Public y la información confidencial ***no*** es ***incluir información confidencial en primer lugar***. Las siguientes no son recomendaciones, sino que simplemente destacan las opciones disponibles.

1. En la discusión sobre la configuración anterior, hay una casilla de verificación para evitar que los usuarios descarguen un libro de trabajo. Es posible que desee seleccionar esa opción donde un libro de trabajo contenga información que no desea que se vea, aparte de lo que elija para hacer visible.
2. Es posible crear una configuración para que un libro de trabajo no se muestre en el perfil de un usuario. Esto es difícil de detectar y aparece al desplazarse sobre el libro de trabajo en la vista de Perfil.

Como el mensaje indica que el uso de esta opción no impide que un libro de trabajo se encuentre en los motores de búsqueda o que los usuarios no vean. Simplemente significa que no está visible en la página de perfil.

Como tal, el público de Tableau se basa fundamentalmente en compartir información con otros a través de la visualización. Es decir, es para información que desea que otros vean. Aquí vale la pena volver brevemente al cuadro de mandos completo y hacer clic en el botón compartir.



Como podemos ver aquí, Tableau genera códigos de inserción para usar en sitios web o para enviar correos electrónicos como un enlace junto con Twitter y Facebook.

9.9 Redondeo

En este capítulo, hemos introducido la visualización de datos de patentes utilizando un conjunto de casi 10,000 documentos de patentes de WIPO Patentscope que mencionan pizza. Como ya debería estar claro, Tableau Public es una herramienta gratuita muy poderosa para la visualización de datos. Requiere atención a los detalles y el cuidado en la construcción, pero es una de las mejores herramientas gratuitas disponibles para la visualización y el tablero.

Para seguir trabajando con Tableau en las patentes de pizza por su cuenta, aquí hay algunos consejos.

1. Ya sabes cómo usar Tableau para crear un mapa de países de publicación.
2. El archivo fuente de pizza contiene un conjunto de números de publicación. Intente a) crear una visualización con los números de publicación, b) buscar en el archivo fuente de la pizza un conjunto de URL y luego explorar qué se puede hacer Worksheet > Actioncon esa URL.
3. En los paneles, considere usar un campo como filtro para otro campo (como el solicitante y el título). ¿Qué fuente de datos o fuentes de datos necesitarías para hacer eso?
4. ¿Qué tipo de historias nos dicen los datos de la pizza y cómo podríamos visualizarlos usando la información provista en los solicitantes y su subconjunto?

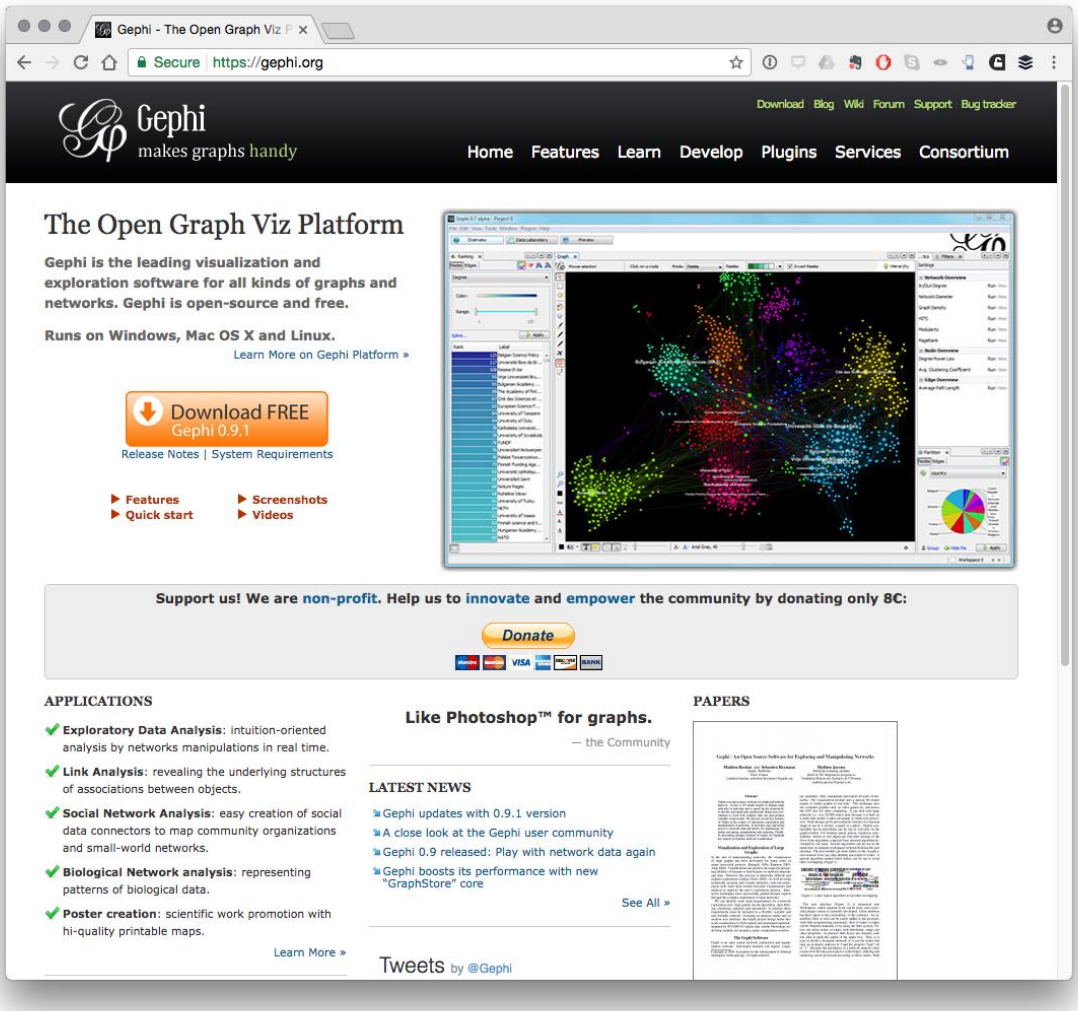
Si se queda atascado y le lleva tiempo familiarizarse con el potencial de Tableau, quizás intente explorar este libro de trabajo sobre biología sintética y el uso de imágenes de Tableau en este artículo de PLOS ONE . Como sugerencia, intente hacer clic en las barras y luego en los títulos para comprender Acciones. Descargar libros de trabajo preparados por otros puede ser una muy buena forma de aprender los consejos y trucos de visualización de cuadros y cuadros de mando.

Si desea descargar el libro de pizza está aquí .

Sin embargo, uno de los problemas más importantes expuestos al trabajar con Tableau es que debe asegurarse de que los campos que desea visualizar sean tidy, que no estén concatenados, y que estén tan limpios como sea razonable para hacerlos. Para los investigadores que deseen elaborar sus propios datos, sugerimos el artículo Open Refine como un buen punto de partida.

Capítulo 10 Gephi

Este capítulo se centra en visualizar datos de patentes en redes utilizando el software de código abierto Gephi . Gephi es uno de un número creciente de herramientas de análisis y visualización de red libre con otros, incluyendo Cytoscape , tulipán , GraphViz , Pajek para Windows, y VOSviewer , por nombrar sólo unos pocos. Además, los paquetes de visualización de red están disponibles para R y Python. Hemos optado por centrarnos en Gephi porque es una buena herramienta de visualización en red que es bastante fácil de usar y de aprender.



En este capítulo nos centraremos en crear una visualización en red simple de la relación entre los solicitantes de patentes (cesionarios). Sin embargo, la visualización de la red se puede utilizar para visualizar un rango de campos y

relaciones, como inventores, palabras clave, códigos de IPC y CPC, y citas, entre otras opciones.

Para este capítulo utilizaremos un conjunto de datos sobre drones de la base de datos de patentes de Lens . El conjunto de datos consta de 5884 documentos de patente que contienen los términos "drone o drones" en el texto completo de duplicado a familias individuales del conjunto completo de publicaciones. El conjunto de datos se ha limpiado de forma exhaustiva en Vantage Point al separar los nombres de los solicitantes e inventores y luego usar la coincidencia lógica difusa para limpiar los nombres. Se pueden lograr resultados muy similares utilizando Open Refine como se describe en el Capítulo 9 de este Manual.

El conjunto de datos se puede descargar desde Github en un archivo zip para descomprimir aquí .

10.1 Instalación de Gephi

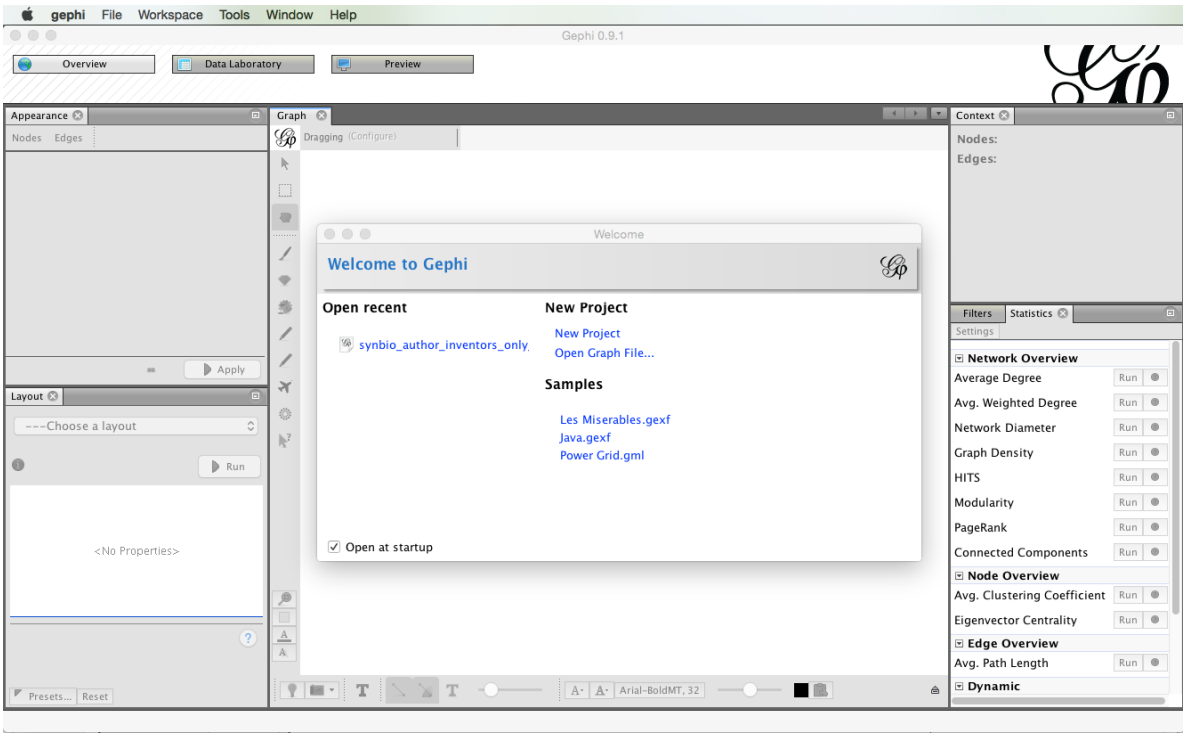
Debe instalar gephi 9.1 (la última versión) en lugar de una versión anterior. Tenga en cuenta que es posible que las actualizaciones posteriores no contengan la funcionalidad clave que se necesita a continuación (ya que algunos de los complementos y funciones tardan un tiempo en recuperarse).

Para instalar para su sistema operativo siga estas instrucciones

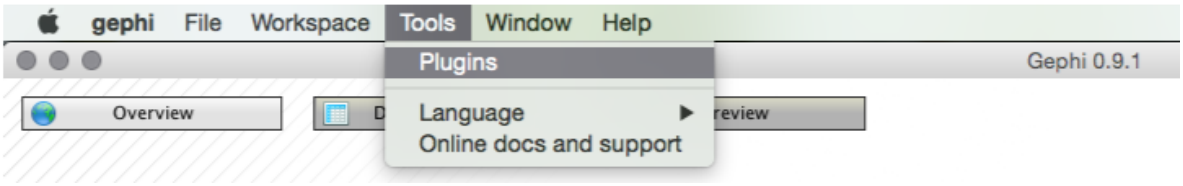
Una vez que haya terminado este capítulo, es posible que desee seguir la guía de inicio rápido, aunque cubriremos esos temas en el artículo. La sección Aprender del sitio web proporciona tutoriales adicionales.

10.2 Apertura de Gephi e instalación de complementos

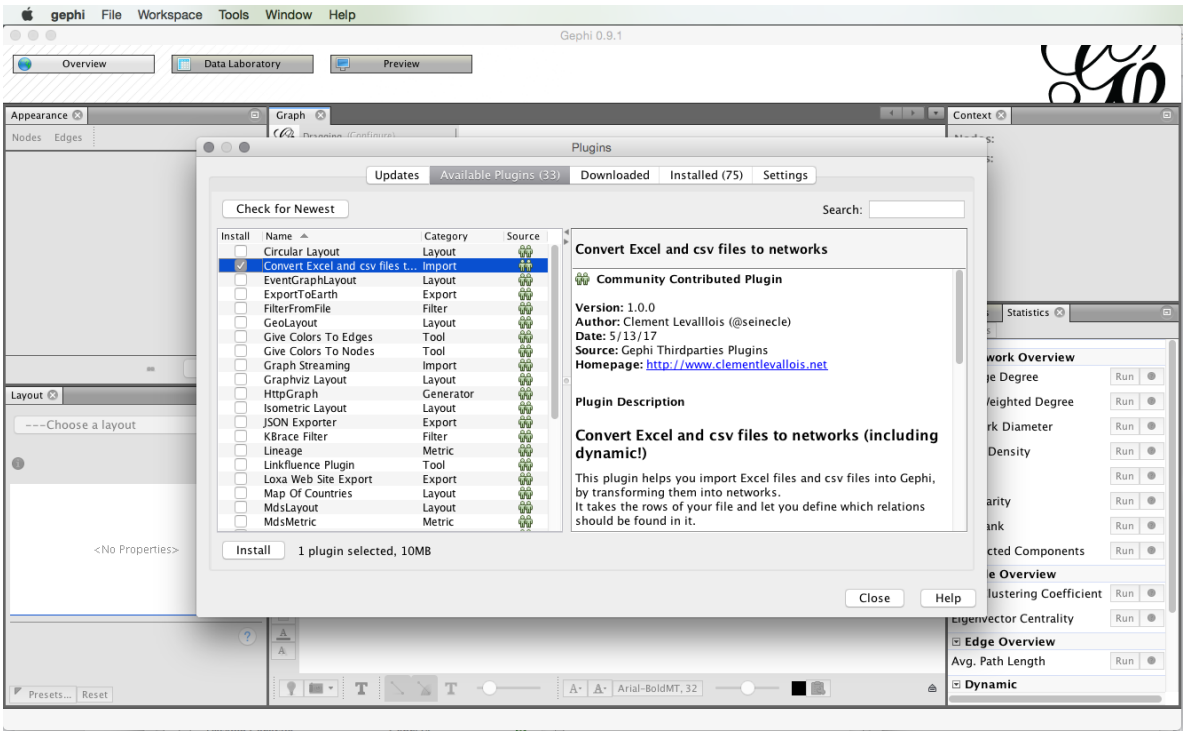
Cuando hayas instalado Gephi, ábrelo y verás la siguiente pantalla de bienvenida.



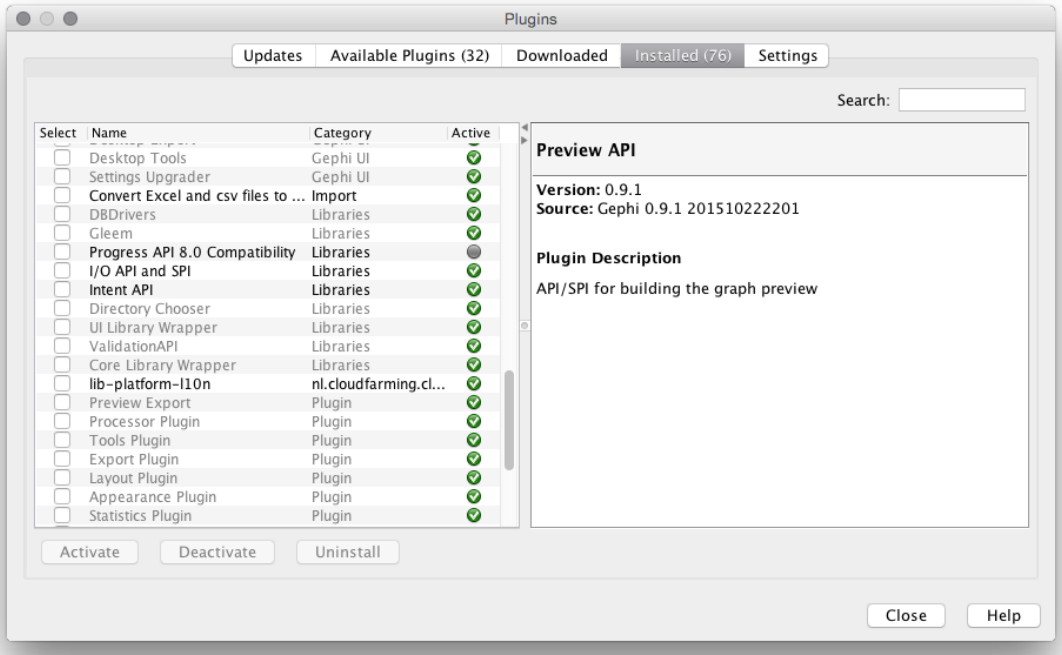
Antes de hacer nada más, necesitamos instalar un complemento desarrollado por Clement Levallois para convertir archivos de Excel y csv en archivos de red gephi. Para instalar el complemento, seleccione el Toolsmenú en la barra de menú y luego Plugins.



Verá un menú emergente para los complementos. En este punto, es posible que desee presionar Reload Catalog para asegurarse de que todo esté cargado. Entonces la cabeza a Available Plugins. Haga clic en name para ordenarlos alfabéticamente. Ahora quieres buscar un plugin llamado Convert Excel and csv files to networks. Seleccione la casilla de verificación, presione Instally siga los menús. Simplemente siga presionando en las indicaciones y luego tendrá que reiniciar al final.



Deberá reiniciar Gephi para que tenga efecto, pero si regresa al menú de Complementos y luego elige la pestaña instalada, debería ver esto.



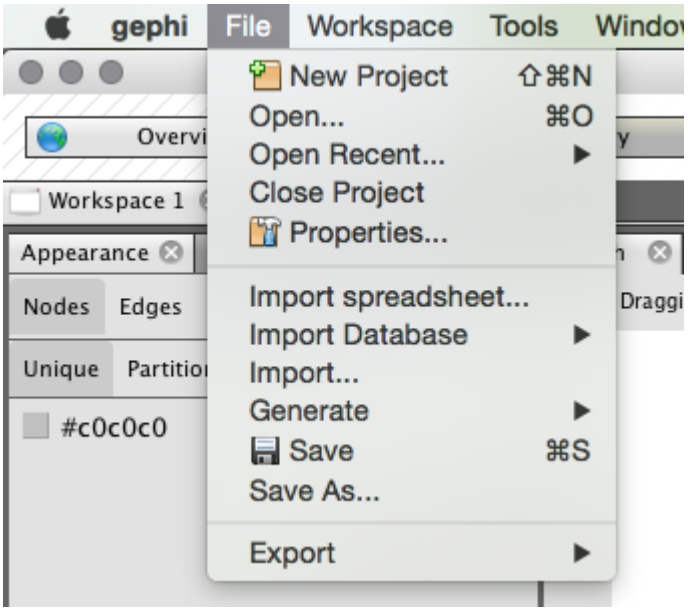
Tú eres bueno para irte. Mientras esté allí, es posible que desee revisar los otros complementos para tener una idea de lo que está disponible. Para obtener más información sobre el complemento de conversión, consulte esta descripción Convertidor de Excel / csv a complemento de red .

10.3 Importando un archivo a Gephi con el plugin convertidor

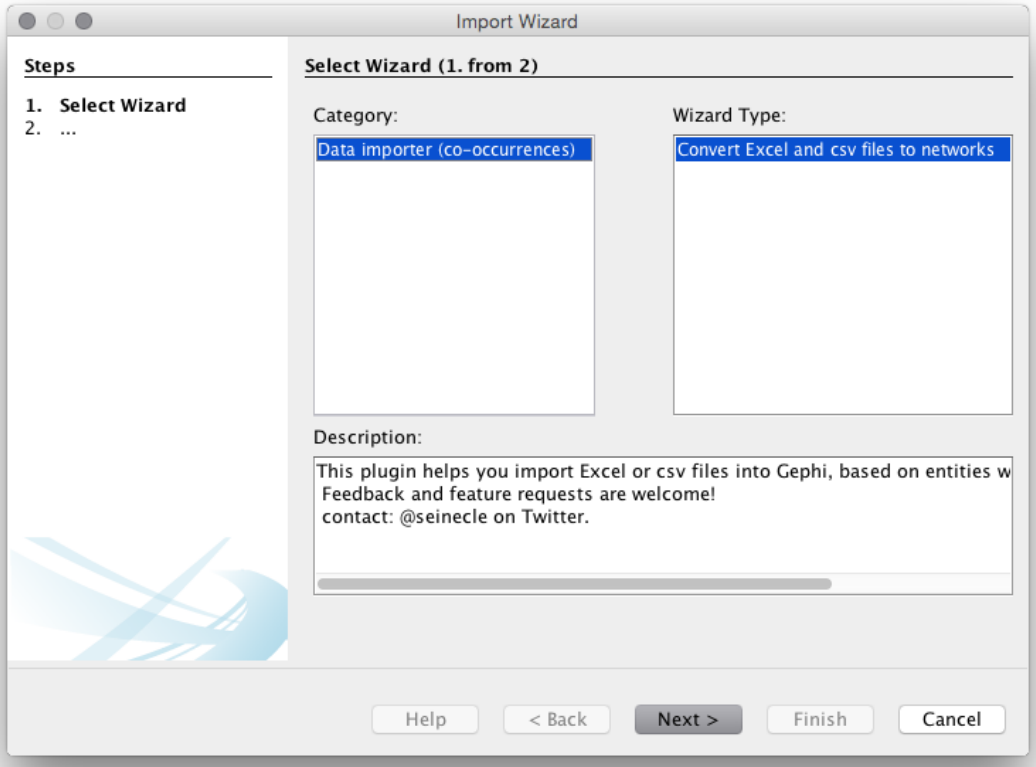
Nos concentraremos en el uso del dron es conjunto de datos de patente en la versión .csv comprimida aquí y no olvide descomprimir el archivo. Mientras Gephi trabaja con archivos .csv, el complemento de importación incluye una opción de línea de tiempo que solo funciona con Excel. Por eso utilizaremos la versión de Excel.

10.3.1 Paso 1. Abra Gephi y elija Archivo> Importar

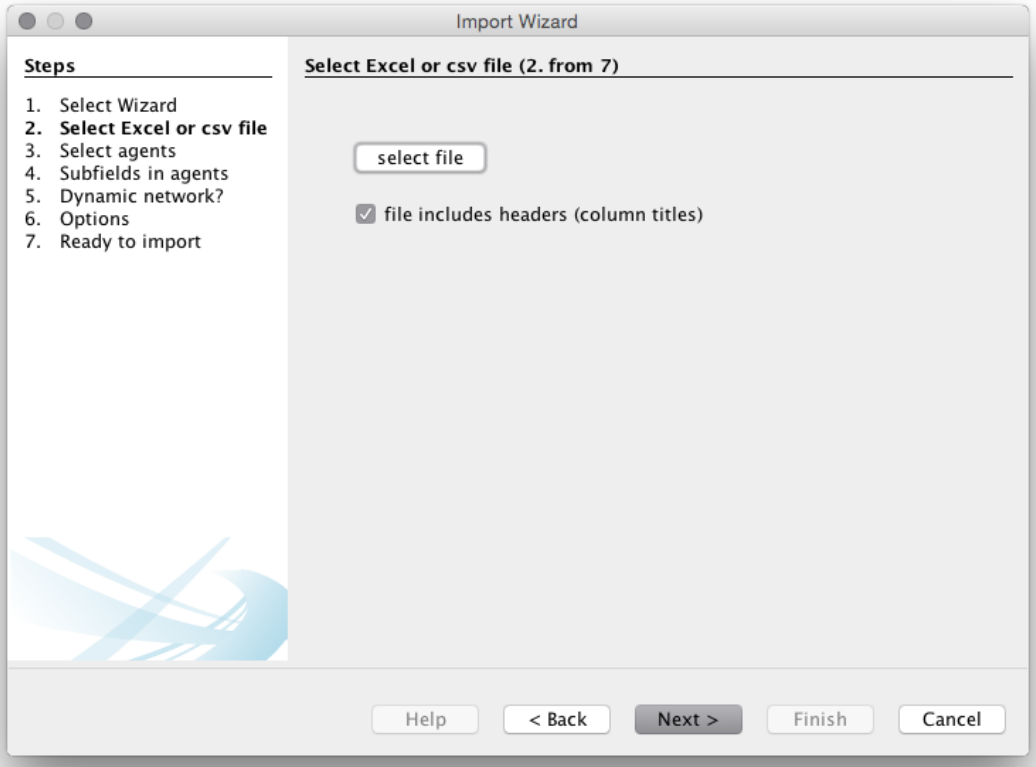
Para que esto funcione, necesitamos usar la Import función en el menú Archivo.



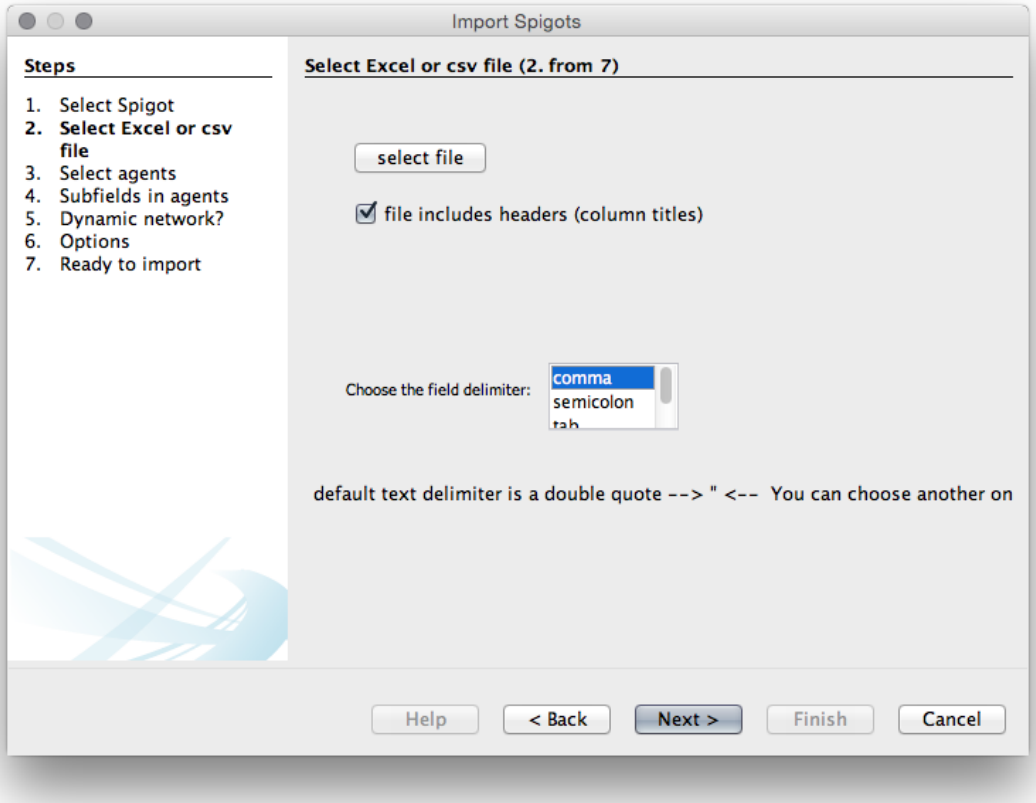
Ahora debería ver un menú como ese a continuación. Asegúrese de que elige la opción de co-ocurrencia.



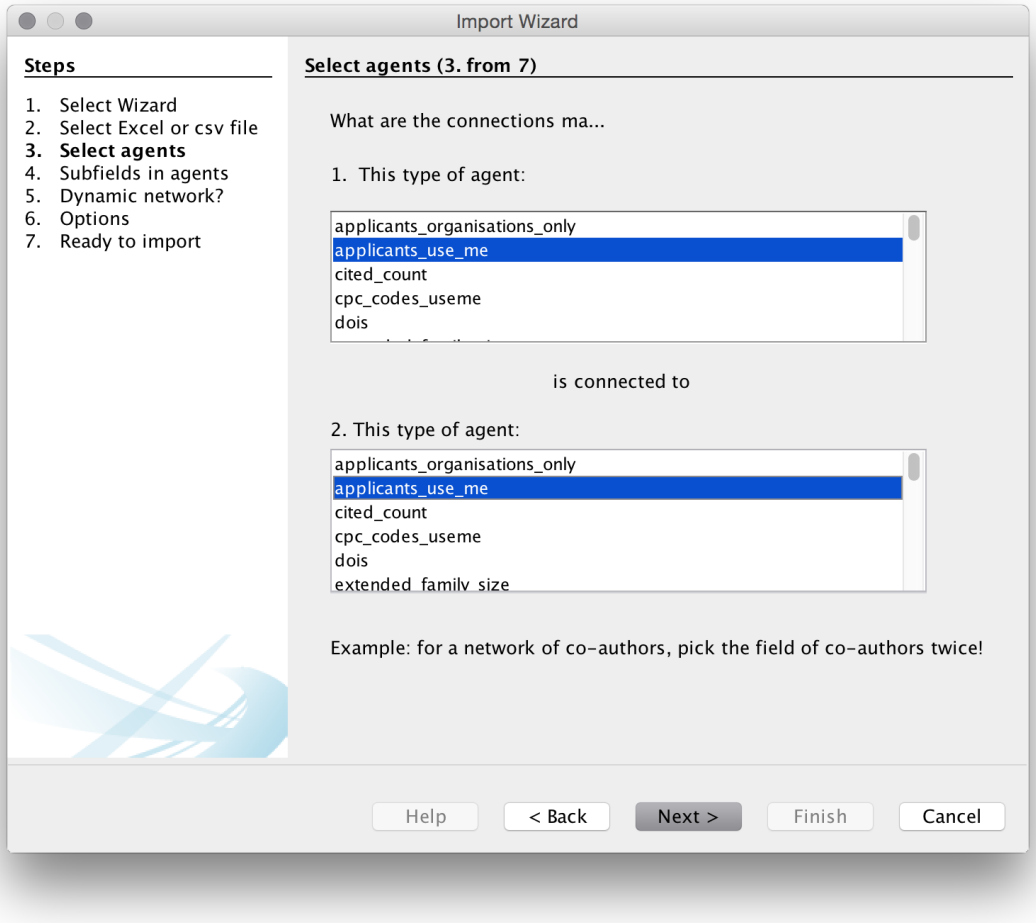
A continuación se le pedirá que seleccione el archivo a utilizar. Descargaremos y luego descomprimiremos el archivo gephi_drones_fulltext_cleaned_5884.csv que se encuentra en el sitio web de WIPO Analytics en Git Hub.



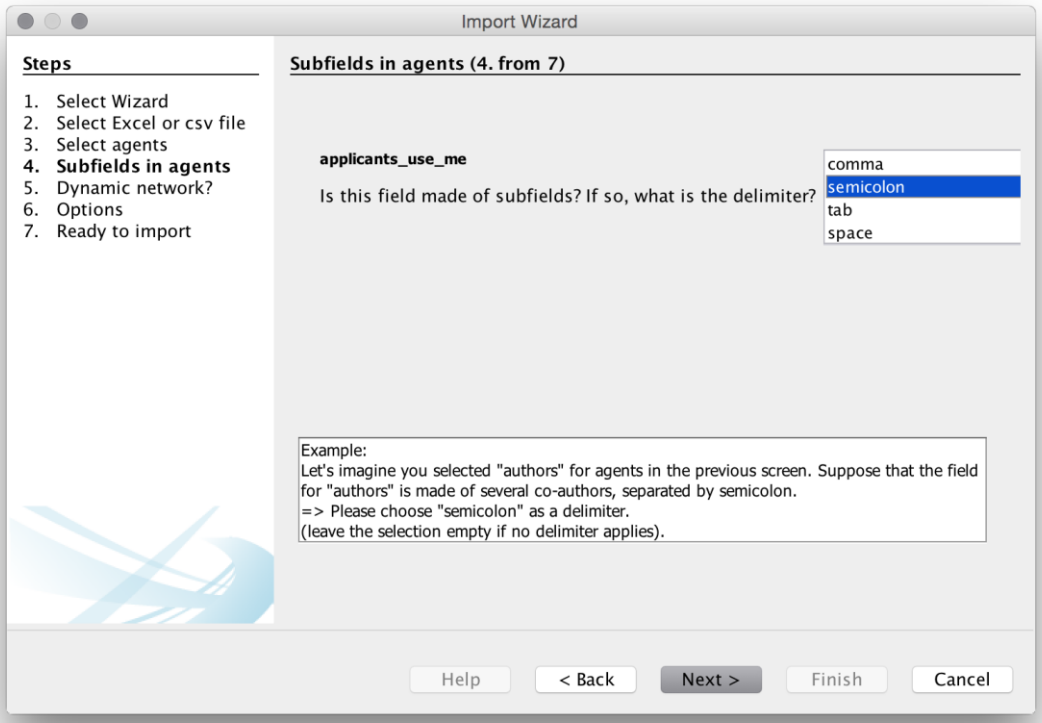
Cuando hayas elegido, Data importer (co-occurrences) entonces elige Next. Asegúrese de que los encabezados de columna permanezcan seleccionados (a menos que use sus propios datos). A continuación, tendrá que elegir un delimitador. En este caso, es una coma, pero en otros casos puede ser un punto y coma o una pestaña.



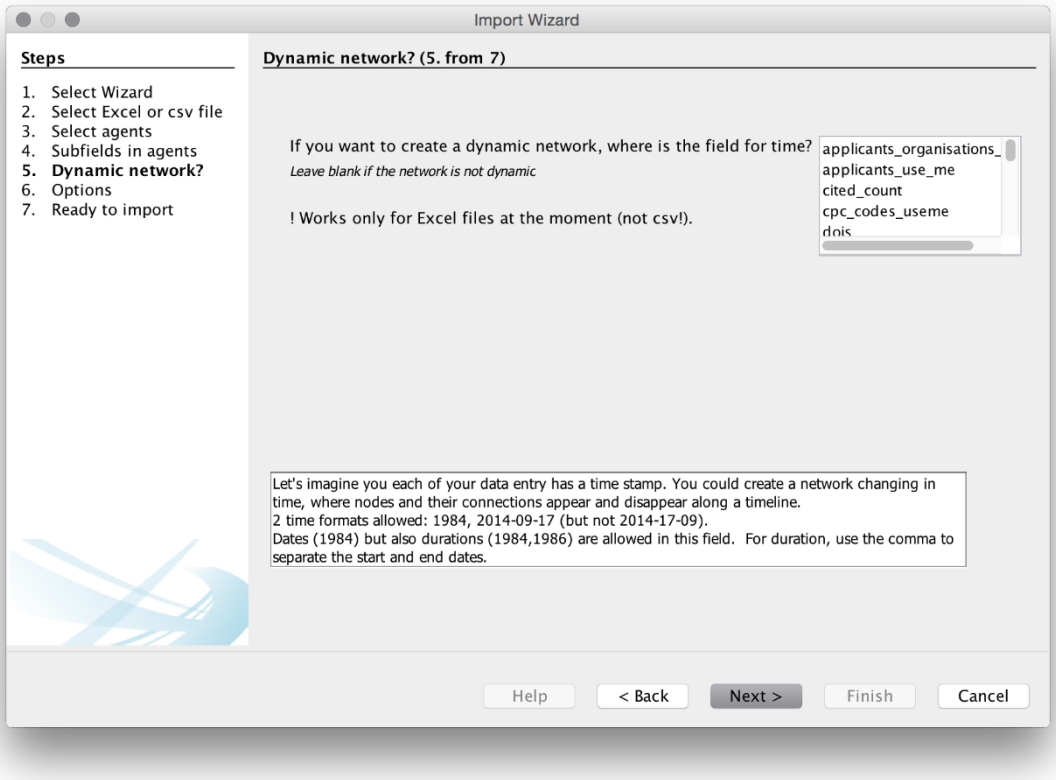
Ahora debemos elegir los agentes, es decir, los actores u objetos con los que queremos crear un mapa de red. Lo usaremos patent_assignees_cleanedya que este es un conjunto relativamente pequeño. Elegiremos el mismo campo en los dos cuadros porque estamos interesados en el análisis de co-ocurrencia.



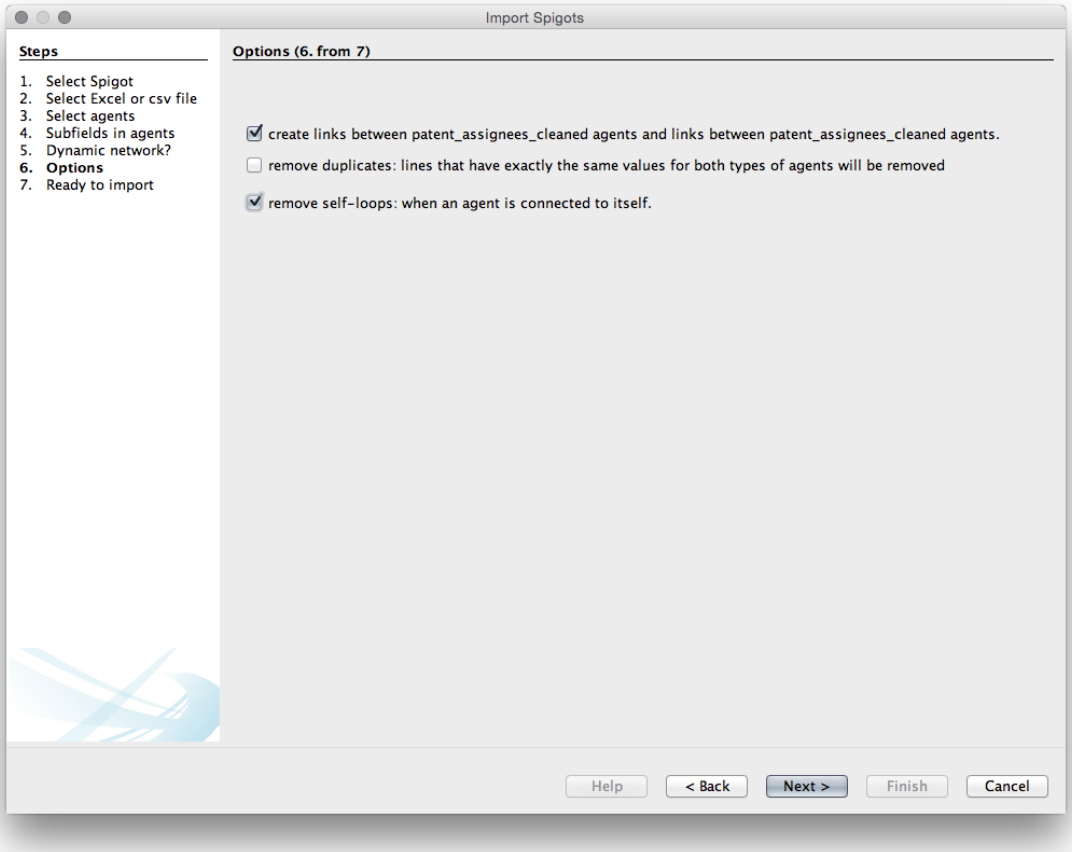
En el siguiente paso, necesitamos especificar el delimitador para dividir el contenido de la applicants_use_mecolumna. En todos los campos es un punto y coma, así que vamos a elegir eso. Tenga en cuenta que, si está haciendo esto con datos de Lentes sin procesar que no ha limpiado previamente, el Límite delimitado es un punto y coma doble (lo que no es útil) y deberá reemplazarse antes de la importación.



Luego se nos preguntará si queremos una red dinámica. Actualmente, esto solo funciona con archivos de Excel e incluso entonces no siempre funciona bien. Dejaremos esto en blanco ya que estamos usando un archivo .csv. Tenga en cuenta que si estuviéramos utilizando un archivo de Excel, las opciones que utilizaríamos normalmente serían el año de publicación o la fecha de publicación o el año o la fecha de prioridad para los datos de patentes.



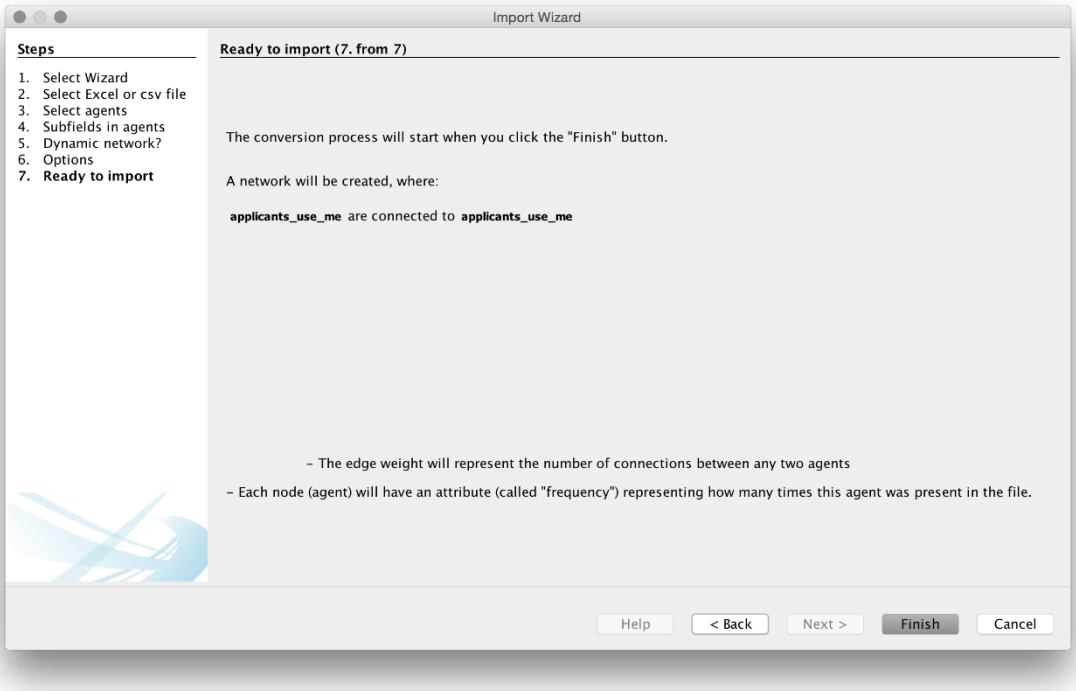
El siguiente menú nos proporciona una lista de opciones. Desafortunadamente, con una excepción, no está del todo claro cuáles son las consecuencias de estas elecciones, por lo que puede ser necesaria la experimentación.



Elección1. Crear enlaces entre applicants_use_me. Opción 2. Eliminar duplicados. No necesitamos eso ya que sabemos que son únicos. Opción 3. Eliminar auto-bucles. Por lo general, queremos esto (también conocido como eliminar la diagonal para evitar que los actores cuenten con ellos mismos, esto producirá un gran aro negro o un asa para un auto-bucle en Gephi).

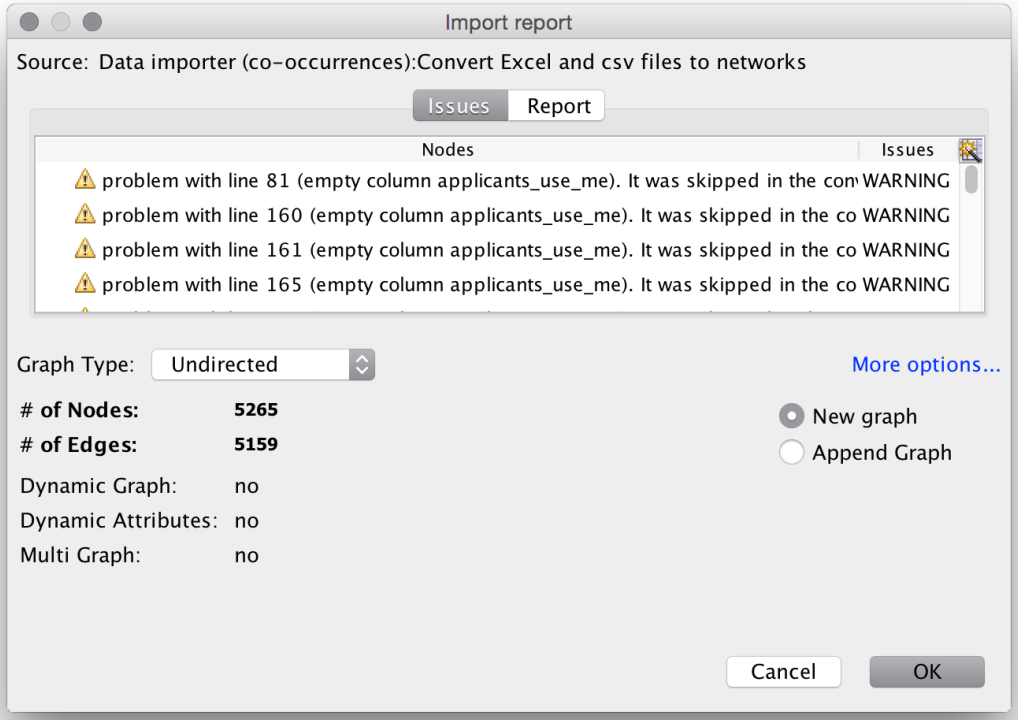
Elegiremos crear los enlaces y eliminar los bucles automáticos.

A continuación, veremos una pantalla de creación de red que establece nuestras opciones.



Presione Finalizar

A continuación, veremos una pantalla de importación.



Es bastante común ver mensajes de advertencia en esta pantalla.

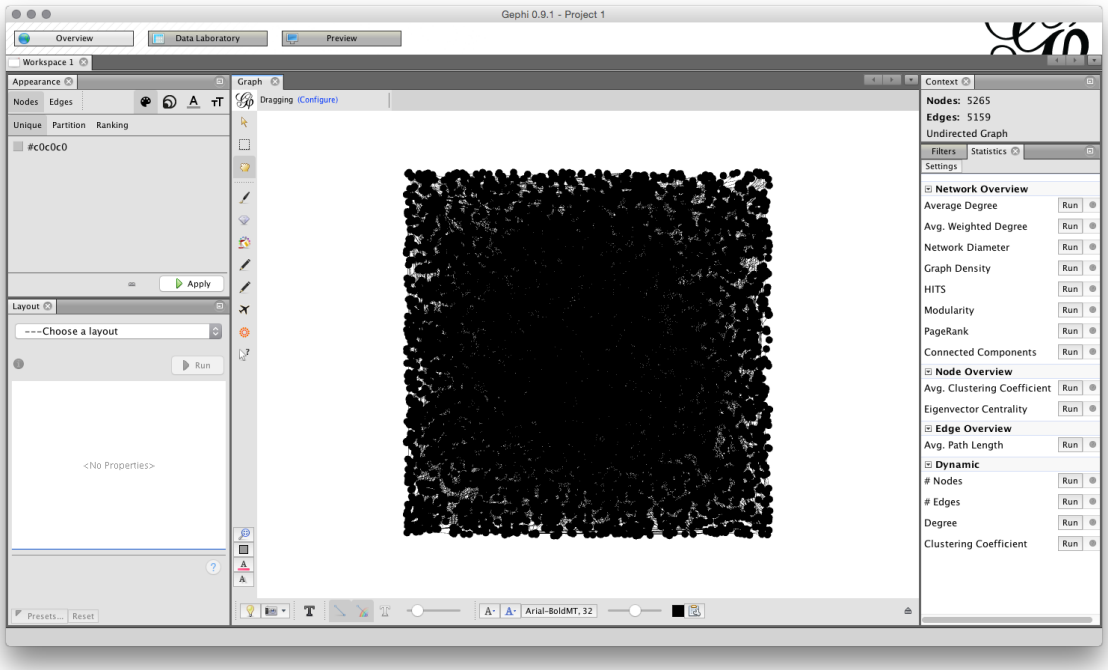
En este caso, algunas de las celdas de los solicitantes en la hoja de trabajo están en blanco porque no hay datos disponibles. Cuando vea mensajes de advertencia, es una buena idea revisar el archivo subyacente para asegurarse de que comprende la naturaleza de la advertencia.

Una segunda advertencia común con las redes dinámicas es que el campo del año no tiene el formato correcto. En ese caso, verifique que el formato del campo de fecha / año sea lo que espera gephi en los datos subyacentes. Puede revisar los datos en el laboratorio de datos.

Tenga en cuenta que la pantalla de importación también proporciona opciones sobre el tipo de gráfico. Normalmente, las redes de autores, inventores y actores dejan esto como una red no dirigida (no ordenada). Undirected es el valor predeterminado básico para los datos de patentes y la literatura científica. También veremos el número de nodos (puntos) y bordes (conexiones). Es importante mantener un ojo en estos valores. Si los nodos son mucho más bajos de lo que

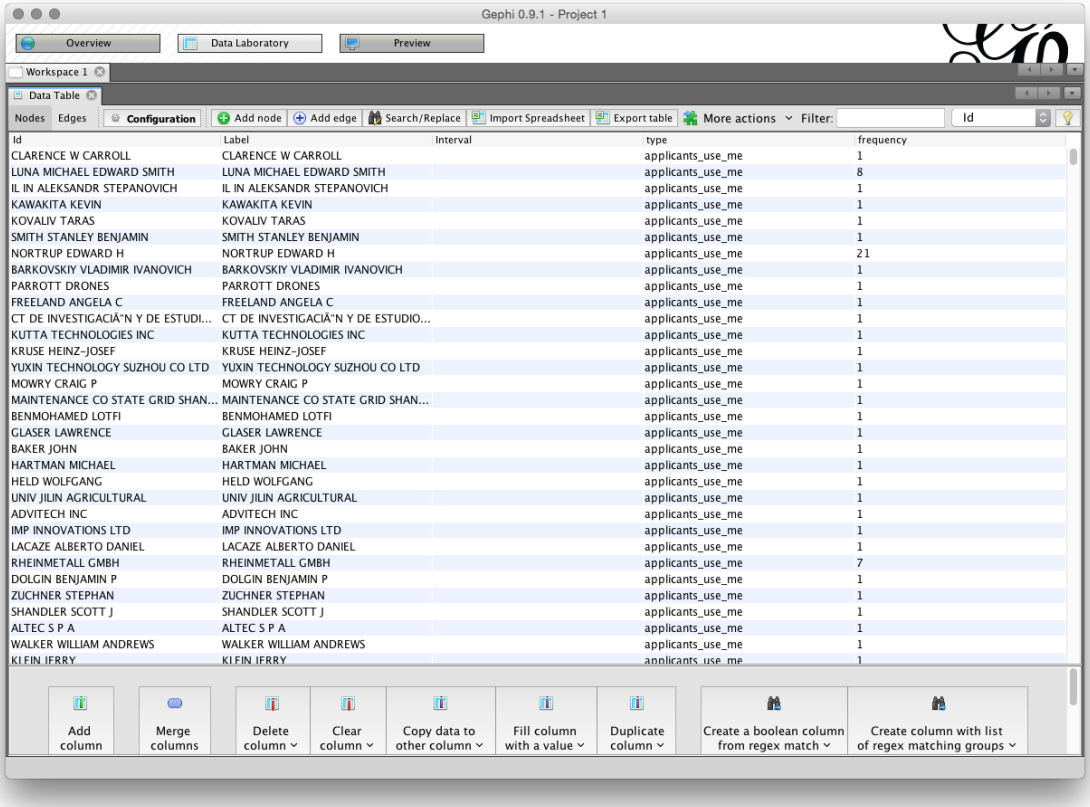
espera, es útil volver a sus datos e inspeccionar problemas como la concatenación de celdas, etc.

Haga clic en Aceptar. Ahora debería ver una red en bruto que se parece a esto.



Tenga en cuenta que podemos ver el número de nodos y bordes en la parte superior derecha. Si cambiamos a la parte superior izquierda, veremos tres pestañas, para Overview, Data Laboratoryy Preview. Elija Data Laboratory.

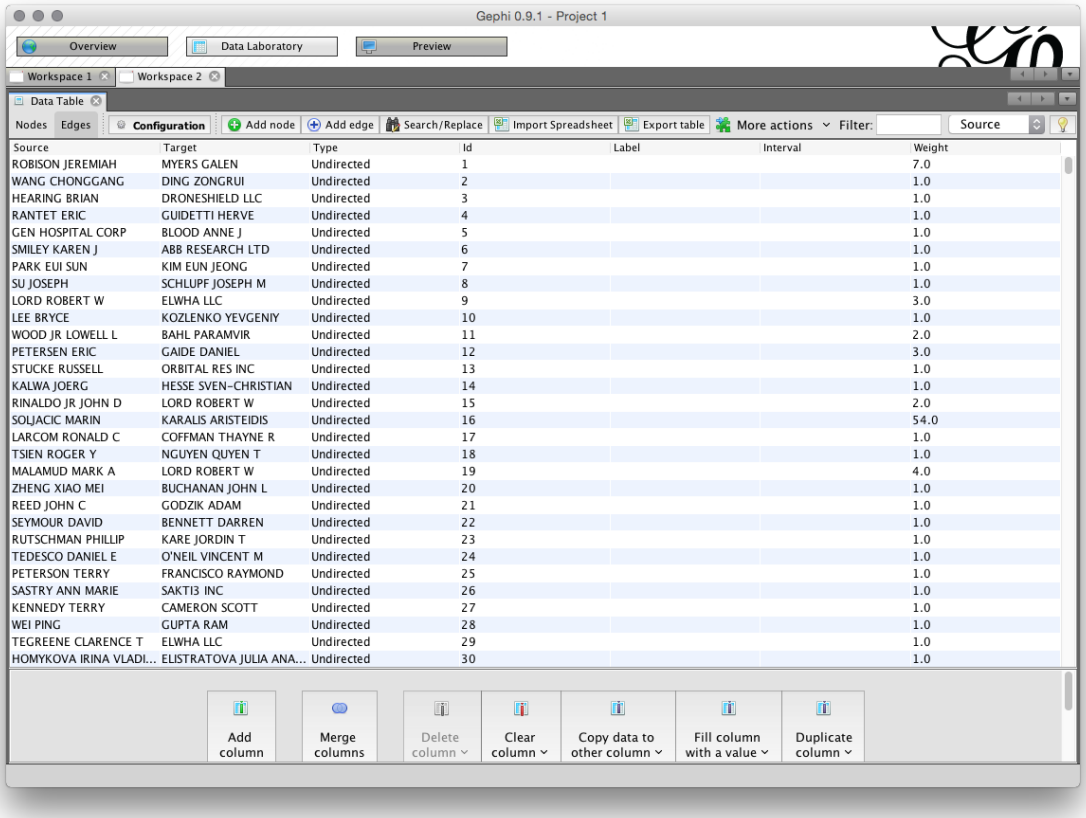
En el laboratorio de datos podemos ver el ID, la etiqueta, el tipo de campo y la frecuencia (el recuento del número de veces que aparece el nombre). Tenga en cuenta que estos campos se pueden editar haciendo clic dentro de la entrada y también se pueden agrupar (por ejemplo, cuando se ha omitido una variante del mismo nombre durante el proceso de limpieza de nombre en una herramienta como Open Refine).



Id	Label	Interval	type	frequency
CLARENCE W CARROLL	CLARENCE W CARROLL		applicants_use_me	1
LUNA MICHAEL EDWARD SMITH	LUNA MICHAEL EDWARD SMITH		applicants_use_me	8
IL IN ALEKSANDR STEPANOVICH	IL IN ALEKSANDR STEPANOVICH		applicants_use_me	1
KAWAKITA KEVIN	KAWAKITA KEVIN		applicants_use_me	1
KOVALIV TARAS	KOVALIV TARAS		applicants_use_me	1
SMITH STANLEY BENJAMIN	SMITH STANLEY BENJAMIN		applicants_use_me	1
NORTRUP EDWARD H	NORTRUP EDWARD H		applicants_use_me	21
BARKOVSKIY VLADIMIR IVANOVICH	BARKOVSKIY VLADIMIR IVANOVICH		applicants_use_me	1
PARROTT DRONES	PARROTT DRONES		applicants_use_me	1
FREELAND ANGELA C	FREELAND ANGELA C		applicants_use_me	1
CT DE INVESTIGACIÃ"N Y DE ESTUDI...	CT DE INVESTIGACIÃ"N Y DE ESTUDIO...		applicants_use_me	1
KUTTA TECHNOLOGIES INC	KUTTA TECHNOLOGIES INC		applicants_use_me	1
KRUSE HEINZ-JOSEF	KRUSE HEINZ-JOSEF		applicants_use_me	1
YUXIN TECHNOLOGY SUZHOU CO LTD	YUXIN TECHNOLOGY SUZHOU CO LTD		applicants_use_me	1
MOWRY CRAIG P	MOWRY CRAIG P		applicants_use_me	1
MAINTENANCE CO STATE GRID SHAN...	MAINTENANCE CO STATE GRID SHAN...		applicants_use_me	1
BENMOHAMED LOTFI	BENMOHAMED LOTFI		applicants_use_me	1
GLASER LAWRENCE	GLASER LAWRENCE		applicants_use_me	1
BAKER JOHN	BAKER JOHN		applicants_use_me	1
HARTMAN MICHAEL	HARTMAN MICHAEL		applicants_use_me	1
HELD WOLFGANG	HELD WOLFGANG		applicants_use_me	1
UNIV JILIN AGRICULTURAL	UNIV JILIN AGRICULTURAL		applicants_use_me	1
ADVITECH INC	ADVITECH INC		applicants_use_me	1
IMP INNOVATIONS LTD	IMP INNOVATIONS LTD		applicants_use_me	1
LACAZE ALBERTO DANIEL	LACAZE ALBERTO DANIEL		applicants_use_me	1
RHEINMETALL GMBH	RHEINMETALL GMBH		applicants_use_me	7
DOLGIN BENJAMIN P	DOLGIN BENJAMIN P		applicants_use_me	1
ZUCHNER STEPHAN	ZUCHNER STEPHAN		applicants_use_me	1
SHANDLER SCOTT J	SHANDLER SCOTT J		applicants_use_me	1
ALTEC S P A	ALTEC S P A		applicants_use_me	1
WALKER WILLIAM ANDREWS	WALKER WILLIAM ANDREWS		applicants_use_me	1
KLEIN JERRY	KLEIN JERRY		applicants_use_me	1

En algunos casos, es posible que haya llenado cualquier celda en blanco del conjunto de datos con NA (para No disponible). Si este es el caso, NA aparecerá como un nodo en la red. Puede abordar este tipo de problema en el Laboratorio de datos haciendo clic derecho en el valor de NA y luego Eliminar. Tenga en cuenta también que siempre puede excluir o combinar nodos después de haber diseñado la red editando en el Laboratorio de datos.

La segunda parte del Laboratorio de datos son los bordes de la Tabla de datos en el Laboratorio de datos. La tabla de bordes incluye un origen y un destino, donde el origen es el nodo de origen y el destino es otro nodo donde hay un enlace entre los nodos. Podemos ver la tabla de bordes ordenada alfabéticamente (haga clic en el encabezado de origen para ordenar) donde el valor en peso es el número de registros compartidos.



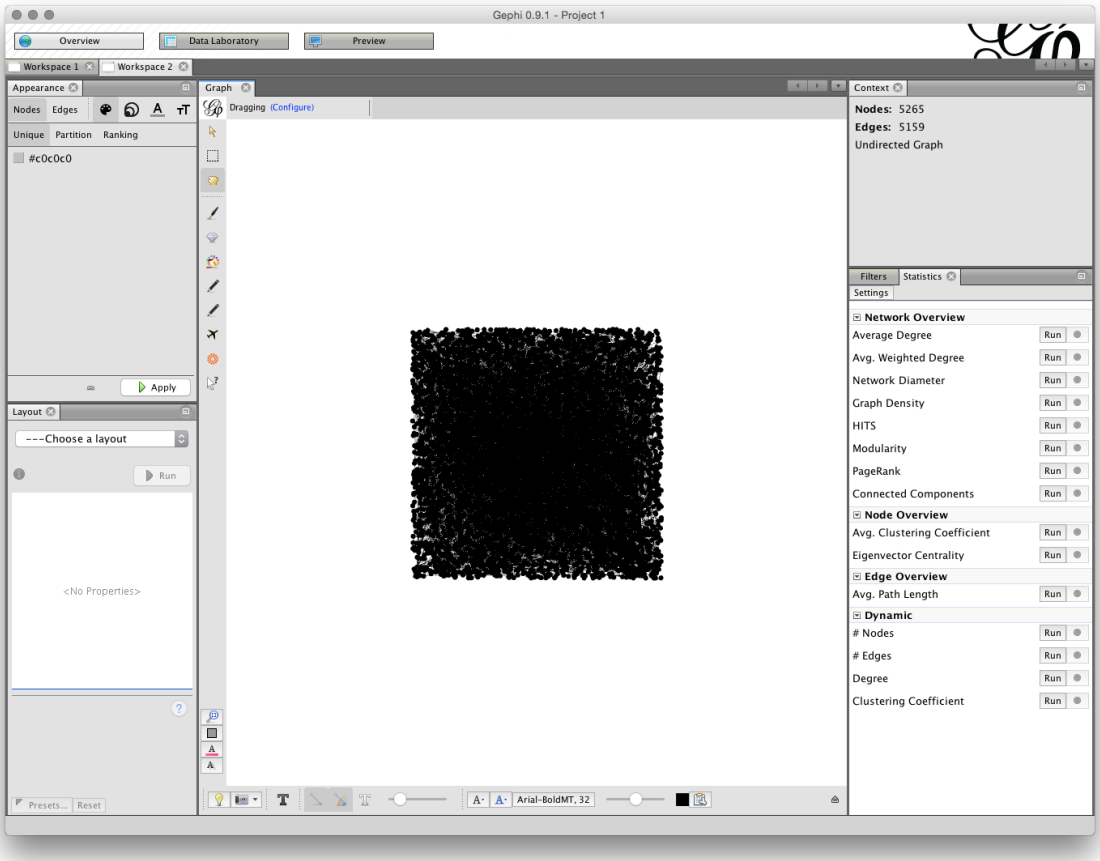
Source	Target	Type	Id	Label	Interval	Weight
ROBISON JEREMIAH	MYERS GALEN	Undirected	1			7.0
WANG CHONGGANG	DING ZONGRUI	Undirected	2			1.0
HEARING BRIAN	DRONESHIELD LLC	Undirected	3			1.0
RANTET ERIC	GUIDETTI HERVE	Undirected	4			1.0
GEN HOSPITAL CORP	BLOOD ANNE J	Undirected	5			1.0
SMILEY KAREN J	ABB RESEARCH LTD	Undirected	6			1.0
PARK EUI SUN	KIM EUN JEONG	Undirected	7			1.0
SU JOSEPH	SCHLUPF JOSEPH M	Undirected	8			1.0
LORD ROBERT W	ELWHA LLC	Undirected	9			3.0
LEE BRYCE	KOZLENKO YEVGENIY	Undirected	10			1.0
WOOD JR LOWELL L	BAHL PARAMVIR	Undirected	11			2.0
PETERSEN ERIC	GAIDE DANIEL	Undirected	12			3.0
STUCKE RUSSELL	ORBITAL RES INC	Undirected	13			1.0
KALWA JOERG	HESSE SVEN-CHRISTIAN	Undirected	14			1.0
RINALDO JR JOHN D	LORD ROBERT W	Undirected	15			2.0
SOLJACIC MARIN	KARALIS ARISTEIDIS	Undirected	16			54.0
LARCOM RONALD C	COFFMAN THAYNE R	Undirected	17			1.0
TSIEN ROGER Y	NGUYEN QUYEN T	Undirected	18			1.0
MALAMUD MARK A	LORD ROBERT W	Undirected	19			4.0
ZHENG XIAO MEI	BUCHANAN JOHN L	Undirected	20			1.0
REED JOHN C	GODZIK ADAM	Undirected	21			1.0
SEYMOUR DAVID	BENNETT DARREN	Undirected	22			1.0
RUTSCHMAN PHILLIP	KARE JORDIN T	Undirected	23			1.0
TEDESCO DANIEL E	O'NEIL VINCENT M	Undirected	24			1.0
PETERSON TERRY	FRANCISCO RAYMOND	Undirected	25			1.0
SASTRY ANN MARIE	SAKTI3 INC	Undirected	26			1.0
KENNEDY TERRY	CAMERON SCOTT	Undirected	27			1.0
WEI PING	GUPTA RAM	Undirected	28			1.0
TEGREENE CLARENCE T	ELWHA LLC	Undirected	29			1.0
HOMYKOVA IRINA VLADI...	ELISTRATOVA JULIA ANA...	Undirected	30			1.0

Nuevamente, tenga en cuenta que es posible exportar el conjunto de bordes e importar un conjunto. También tenga en cuenta los menús en la parte inferior de la pantalla que permiten copiar los valores de las columnas. Esto puede ser útil cuando el valor de la etiqueta no se rellena, lo que significa que un nombre no se mostrará en el nodo cuando se distribuya el gráfico.

La mayoría de las veces simplemente podemos proceder con el diseño de la red sin prestar mucha atención al laboratorio de datos. Sin embargo, es importante familiarizarse con el laboratorio de datos para evitar problemas inesperados o detectar la corrupción en los datos.

10.4 Nodos de dimensionamiento y coloración

Cuando miramos la Overview pantalla tenemos una amplia gama de opciones. Comenzaremos en la parte superior derecha con el panel de estadísticas.



Los Run botones calcularán un rango de estadísticas en la red. Probablemente los dos más útiles son:

1. Clase de modularidad. Este algoritmo recorre las conexiones (bordes) y asigna los nodos a comunidades o agrupaciones en función de la fuerza de las conexiones. Este algoritmo se explica en detalle en este artículo Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre, despliegue rápido de comunidades en grandes redes, en Journal of Statistical Mechanics: Theory and Experiment 2008 (10), P1000 . La capacidad de detectar comunidades en redes basadas en la fuerza de las conexiones es una herramienta poderosa en el análisis de patentes.
2. Diámetro de la red. Esto calcula dos medidas de betweeness, es decir betweeness centrality(la frecuencia con la que aparece un nodo en la ruta más corta entre nodos) y la centralidad (la distancia promedio desde un nodo inicial a otros nodos en la red). El diámetro de la red también calcula la excentricidad, que es la distancia entre un nodo determinado y el nodo más

lejano desde la red. Para obtener información sobre esto, consulte la entrada de Wikipedia y también Ulrik Brandes, un algoritmo más rápido para la centralidad de la intermediación, en Journal of Mathematical Sociology 25 (2): 163-177, (2001)

Mientras que la clase de modularidad identifica comunidades (particularmente en redes grandes), las medidas de centralidad examinan la posición de un nodo en el gráfico en relación con otros nodos. Esto puede ser útil para identificar actores clave en redes según la naturaleza de sus conexiones con otros actores (en lugar de simplemente el número de registros).

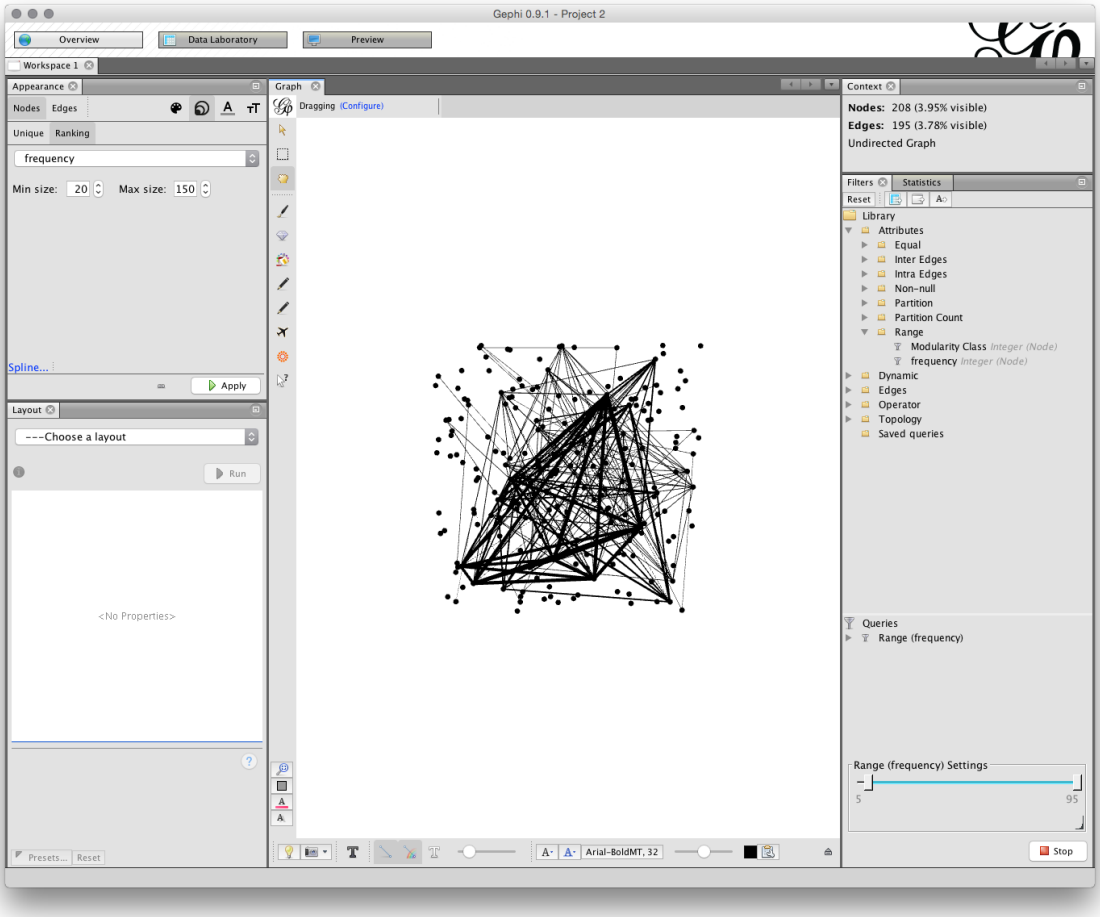
Si ejecutamos Modularity Class como en la figura, un mensaje emergente nos informará que hay 246 comunidades en la red. Dado que solo hay 362 nodos, esto sugiere una red débilmente conectada formada por pequeños grupos individuales.

10.4.1. Filtrar los datos.

Tenemos un total de 5,265 nodos que es bastante denso. Después de ejecutar el algoritmo de clase de modularidad anterior, ahora pasaremos a la pestaña Filtros junto a Estadísticas. Nuestro objetivo aquí es reducir el tamaño de la red.

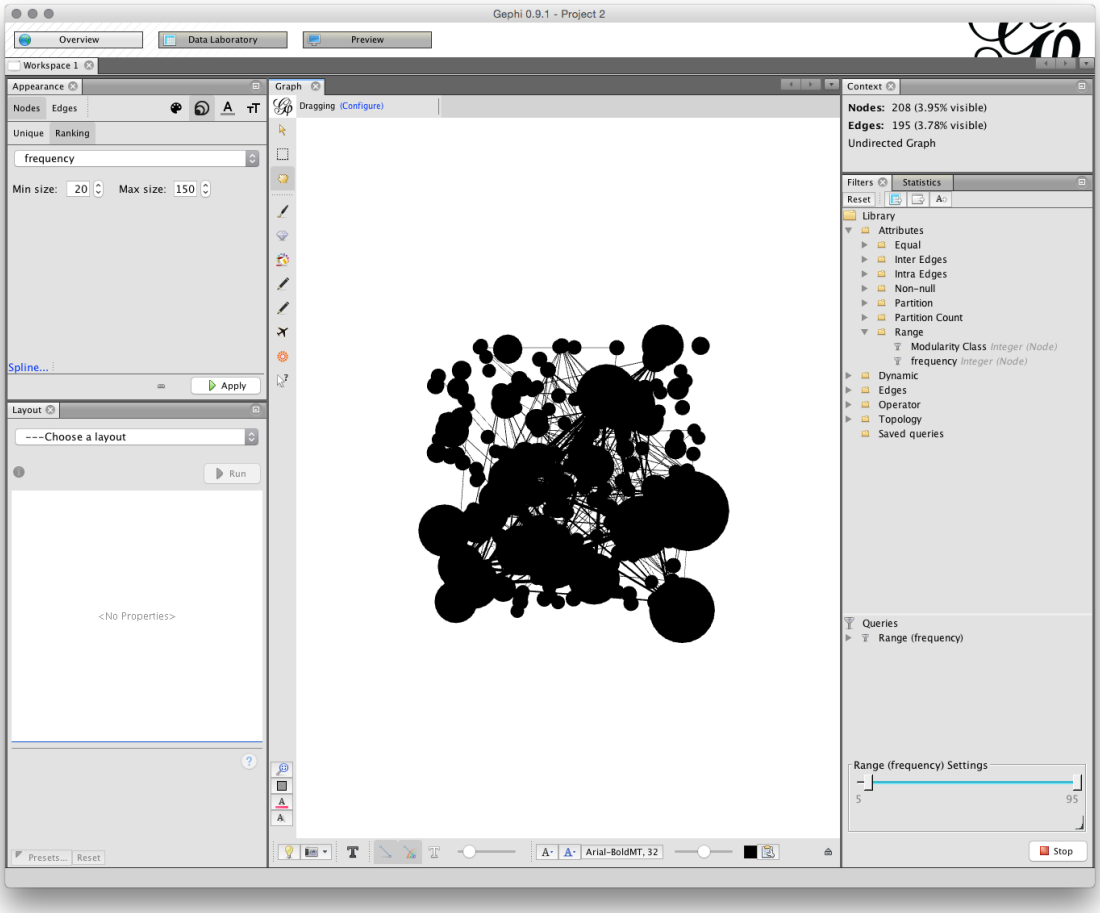
Muévase hacia la izquierda donde dice Clasificación y luego seleccione el triángulo invertido rojo. Establezca el valor más grande en 200 y el más pequeño en 20 (depende de usted lo que elija). Entonces aplique. La red ahora cambiará.

Abra el menú Filtros y elija Atributos. Eso abrirá un conjunto de carpetas y nos gustaría usar el rango. Cuando la carpeta de Rango está abierta, arrastre la frecuencia al área de Consultas a continuación (marcada con un icono rojo y arrastre el mensaje cuando esté vacía). Luego, arrastre la barra de rango hasta que vea una frecuencia de 5 como mínimo o cambie el número haciendo clic en ella. Tenga en cuenta que a medida que arrastramos los resultados, la cantidad de Nodos y Bordes en el Contexto anterior cambiará. Estamos buscando un número manejable. En la imagen de abajo he puesto el número a 5.



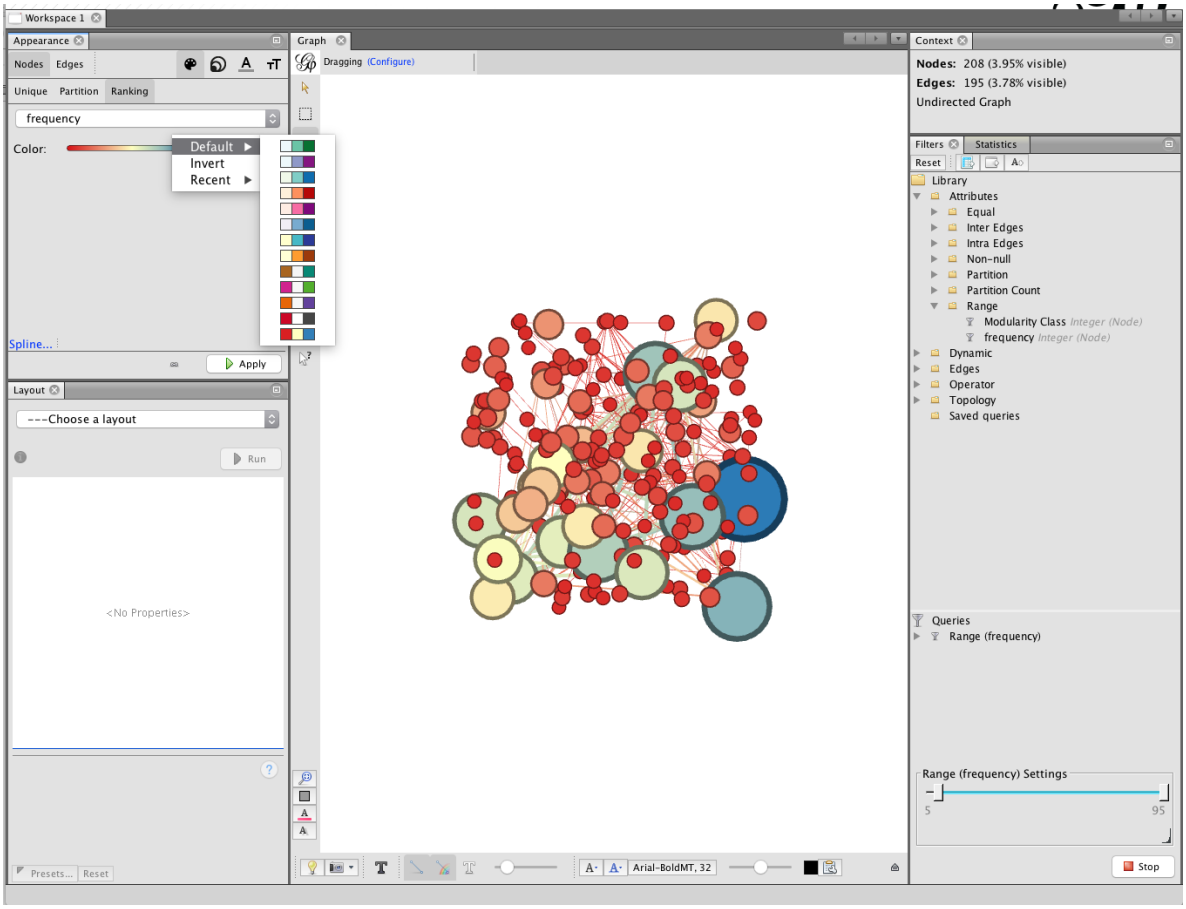
10.4.2 Configuración del tamaño del nodo

A continuación, queremos dimensionar los nodos. A la izquierda, busque la pestaña Apariencia y luego, con Nodos en gris, elija el botón Clasificación. Aquí, el tamaño mínimo se establece en 20 y el máximo en 150. Tenga en cuenta que la configuración predeterminada es 10 y esto generalmente es demasiado pequeño para una fácil visibilidad. Presione Aplicar y verá los cambios en la red para mostrar el tamaño de los nodos según la frecuencia. Siempre puede ajustar el tamaño de los nodos más tarde si no está satisfecho con ellos.



10.4.3 Coloreando los Nodos

Para colorear los nodos, elija el icono de paleta pequeña al lado del icono de tamaño. Ahora tenemos opciones en Único (simplemente gris), Partición o Clasificación. En este caso elegiremos Clasificación y frecuencia. Tenga en cuenta que se puede acceder a una gama de paletas de colores haciendo clic en el pequeño icono a la derecha de la barra de colores debajo de la clasificación. Cuando haya encontrado una paleta que le guste, haga clic en Aplicar.



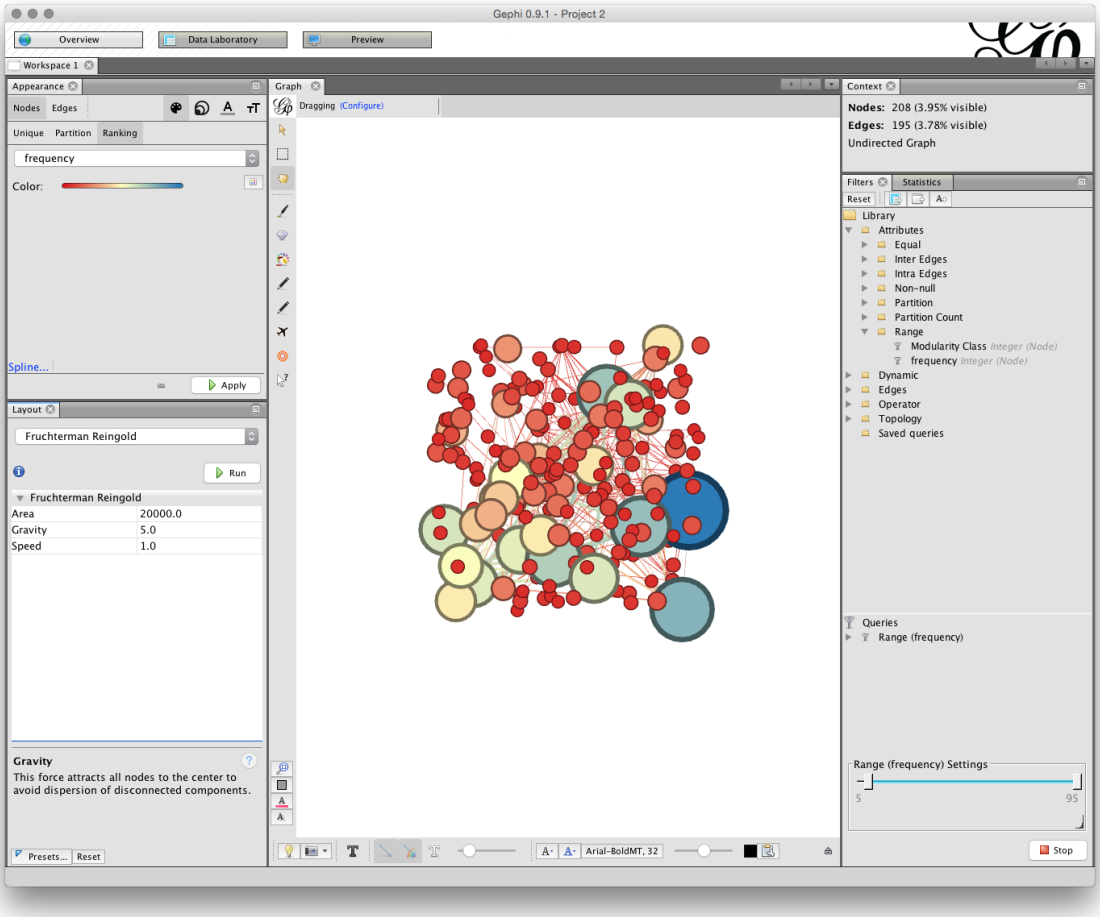
Una forma alternativa de colorear el gráfico en versiones anteriores de gephi era particionar en la clase de Modularidad. Esto colorearía los nodos como "comunidades" de nodos estrechamente vinculados. Sin embargo, en la actualidad, en Gephi 9, esta opción no parece estar disponible de manera constante, pero puede volver en una actualización futura.

Hay una gama de otras opciones para colorear nodos, incluido un complemento de color que detectará si una columna con un valor de color está presente en los datos importados. Esto puede ser muy útil si tiene categorías de datos codificados por colores antes de importar a gephi.

10.5 Diseño del gráfico

En el panel inferior izquierdo llamado Diseño, en la figura anterior, hay una gama de opciones de visualización de red con más que se pueden importar desde los menús de los complementos. Entre los más útiles se encuentran Fruchterman-Reingold, Force Atlas, OpenOrd y Circular con complementos especializados para diseños georreferenciados con los que vale la pena experimentar.

Ilustraremos el diseño de la red utilizando Fruchterman-Reingold. El primer ajuste es el área para el gráfico. El valor predeterminado es 10,000, pero comenzaremos con 20,000 porque 10,000 tiende a estar demasiado aplastado. El valor predeterminado para la configuración de la gravedad es 10. Está pensado para evitar que los nodos se dispersen demasiado, pero a menudo es demasiado apretado cuando se aplican las etiquetas. Intente cambiar la configuración a 5 (lo que reduce el tirón gravitacional). Las configuraciones en las diferentes opciones de diseño pueden tardar un poco en acostumbrarse y vale la pena crear un registro de configuraciones útiles. Gephi no guarda sus configuraciones, así que asegúrese de anotar las configuraciones útiles.



Ahora estamos bien para ir Pero, antes de comenzar, tome nota de dos opciones importantes para su uso posterior.

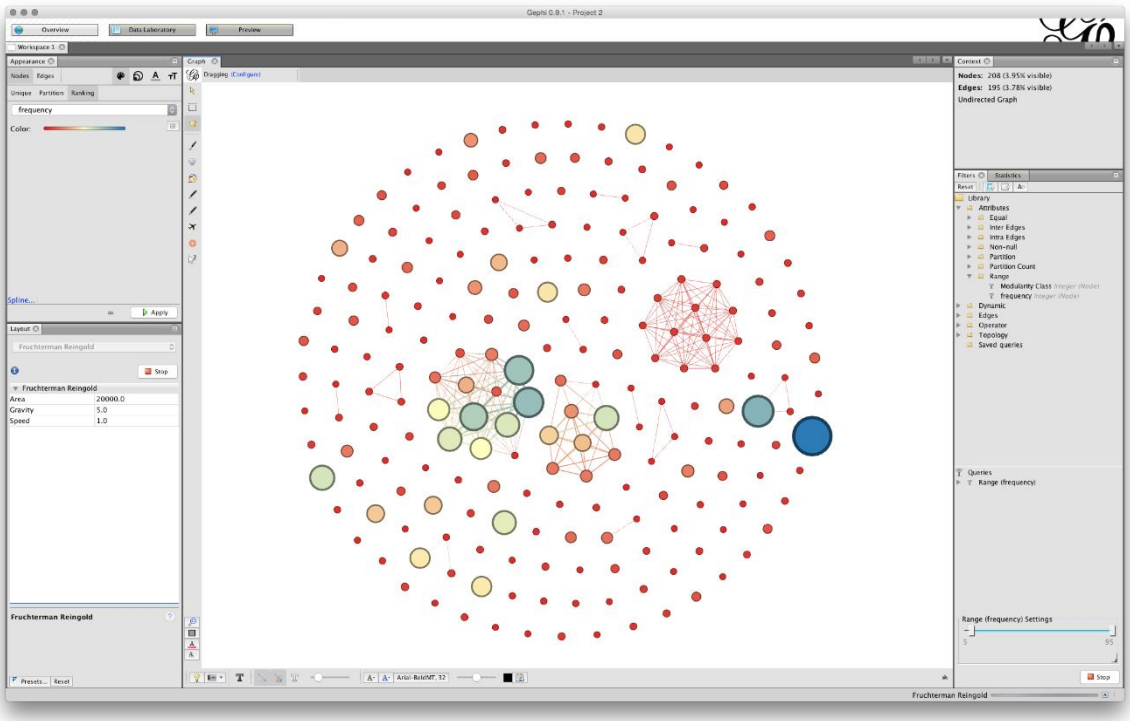
El primero es el complemento NoOverlap que instalamos anteriormente. Esto nos ayudará a lidiar con los nodos traslapados después del diseño. La segunda es la expansión, que nos ayudará a aumentar el tamaño de una red para que sea más fácil

ver las etiquetas. Tenga en cuenta también la opción de Contracción que nos permitirá volver a conectar una red si se expande demasiado.

Ahora asegúrese de que Fruchterman-Reingold esté seleccionado con la configuración mencionada anteriormente y haga clic Run.

Puede dejar que la red se ejecute y los nodos comenzarán a asentarse. Si la red desaparece de la vista (según el mouse), intente desplazarse para alejar la imagen. Nuestro objetivo es llegar a una situación en la que las líneas solo crucen a través de los nodos donde están conectadas. A medida que adquiera más experiencia con el diseño, es posible que desee ayudar a los nodos a moverse hacia una posición clara para obtener un gráfico más ordenado.

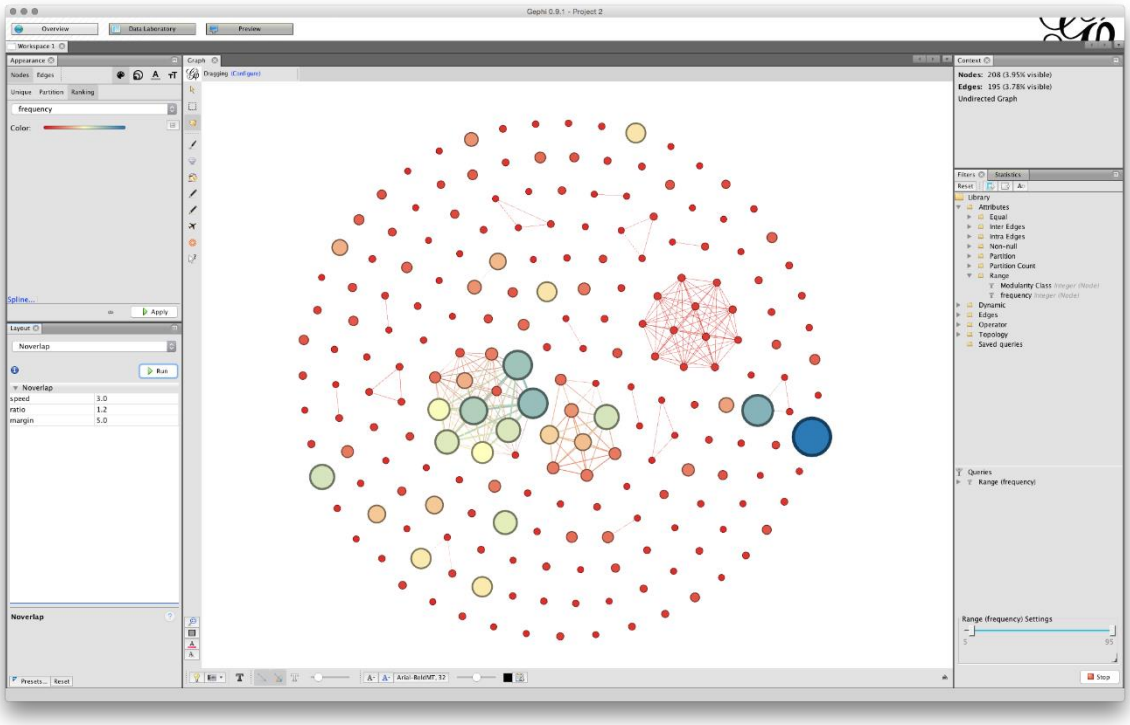
Ahora tendrá una red que se verá así (tenga en cuenta que 15,000 para el Área pueden haber sido suficientes).



Podemos ver que algunos de los nodos están muy juntos. Eso afectará la capacidad de etiquetar los nodos de una manera clara. Para solucionar esto, primero usamos la función NoOverlap y luego deseamos usar la función Expansión en los elementos del menú desplegable Diseño.

Elija nooverlap en el menú y Ejecutar.

Si bien la diferencia es muy pequeña en este caso, al menos hemos movido los nodos a posiciones separadas. En una etapa posterior es posible que desee utilizar la función Expansión. Esto aumentará el tamaño de la red y es útil cuando se trabaja con etiquetas.



10.5.1 Guarda tu trabajo

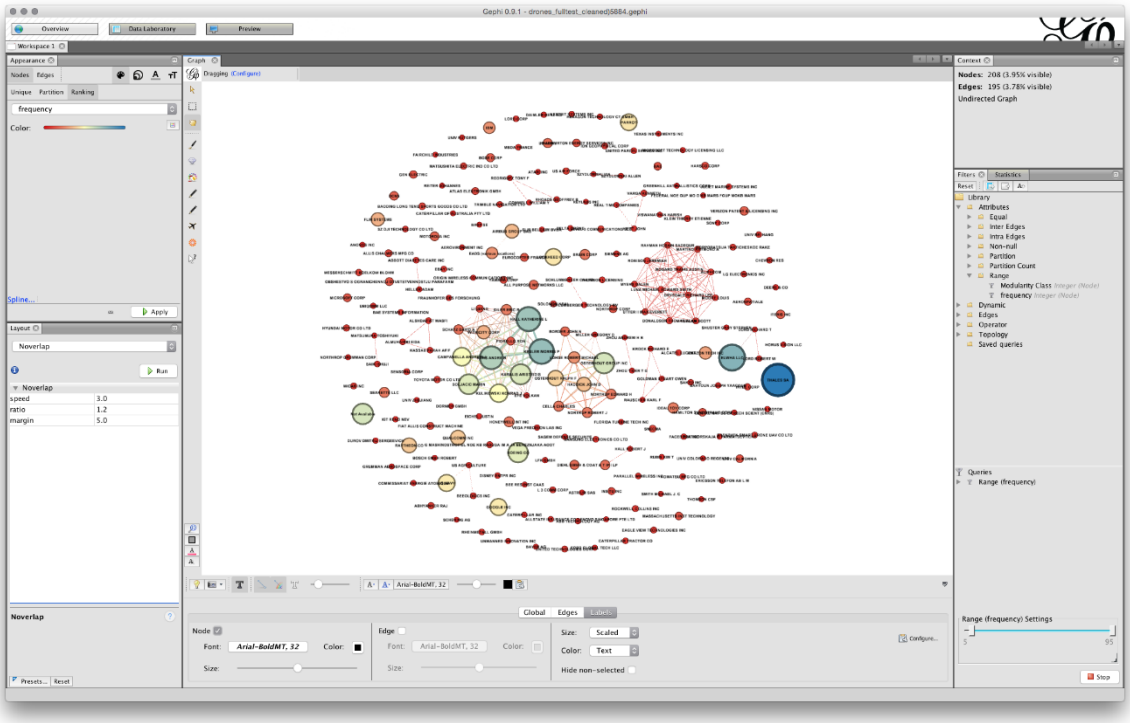
En esta etapa salvaremos nuestro trabajo. Una característica de Gephi como programa Java es que no hay opción de deshacer. Como resultado, es una buena idea guardar el trabajo en un punto en el que esté bastante satisfecho con el diseño tal como está.

Vaya a Archivo y elija Guardar como y asigne una .gephiextensión al nombre del archivo . No olvides hacer esto o gephi no sabrá cómo leer el archivo. Si todo va bien el archivo se guardará. En algunas ocasiones, Java puede lanzar una excepción y, básicamente, tendrá que comenzar de nuevo. Esa es una razón para ahorrar trabajo en Gephi con regularidad porque es un programa beta y está sujeto a las predilecciones de Java en su computadora.

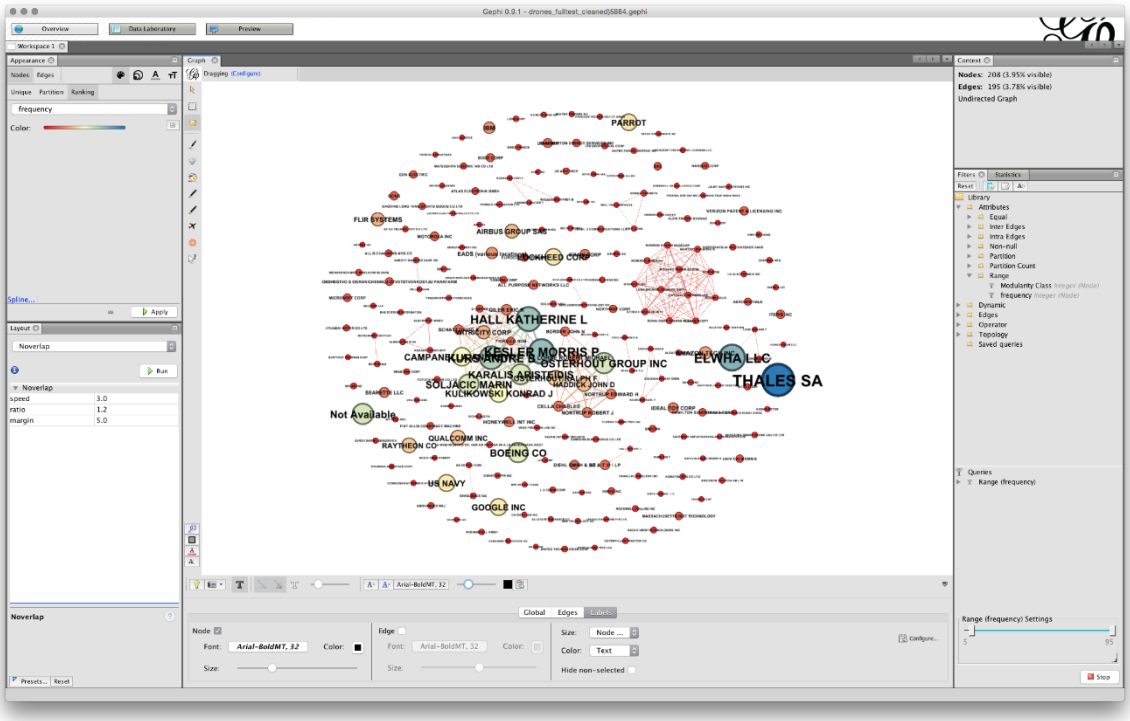
10.6 Adición de etiquetas

El siguiente paso es agregar algunas etiquetas. En la barra de menú inferior hay una gama de opciones. Lo que queremos es el pequeño triángulo gris a la derecha de esta barra de menú que abrirá una nueva barra. Haga clic en el triángulo y verá un conjunto de opciones. Elija las etiquetas y luego en el extremo izquierdo marque la Nodecasilla. No veremos ninguna etiqueta todavía.

A la derecha hay un menú con tamaño. Esto se establece a escala. Para ver algunas etiquetas, mueva el control deslizante de escala hasta el tope. Veremos las etiquetas a la vista y un primer indicio de que tendremos que trabajar un poco más en la disposición del gráfico para que sea legible.



A continuación, cambie el tamaño a Tamaño del nodo, la pantalla se llenará de texto. Vaya al escalador y tire de él hacia atrás hasta que haya algo más o menos legible.

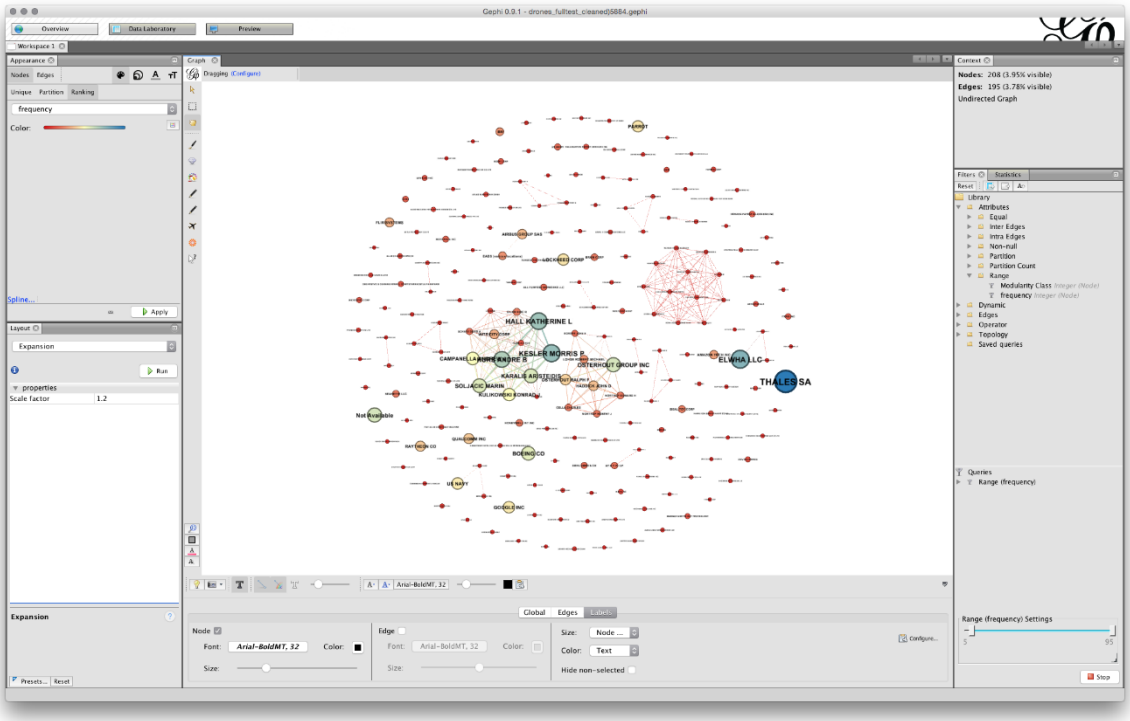


En esta etapa es posible que tengamos que tomar un par de acciones.

1. Cuando quede claro que nuestros nodos están demasiado juntos, tendremos que ejecutar Expansión desde el menú de diseño. Como regla general, solo deberías hacer esto dos veces como máximo ... pero puede depender de tu gráfica.
2. Si tiene etiquetas muy largas, como el Instituto de Tecnología de Massachusetts, probablemente querrá dirigirse al Observatorio y editar la Etiqueta del nodo para que sea manejable, como MIT. Esto puede hacer una gran diferencia en la limpieza de las etiquetas.

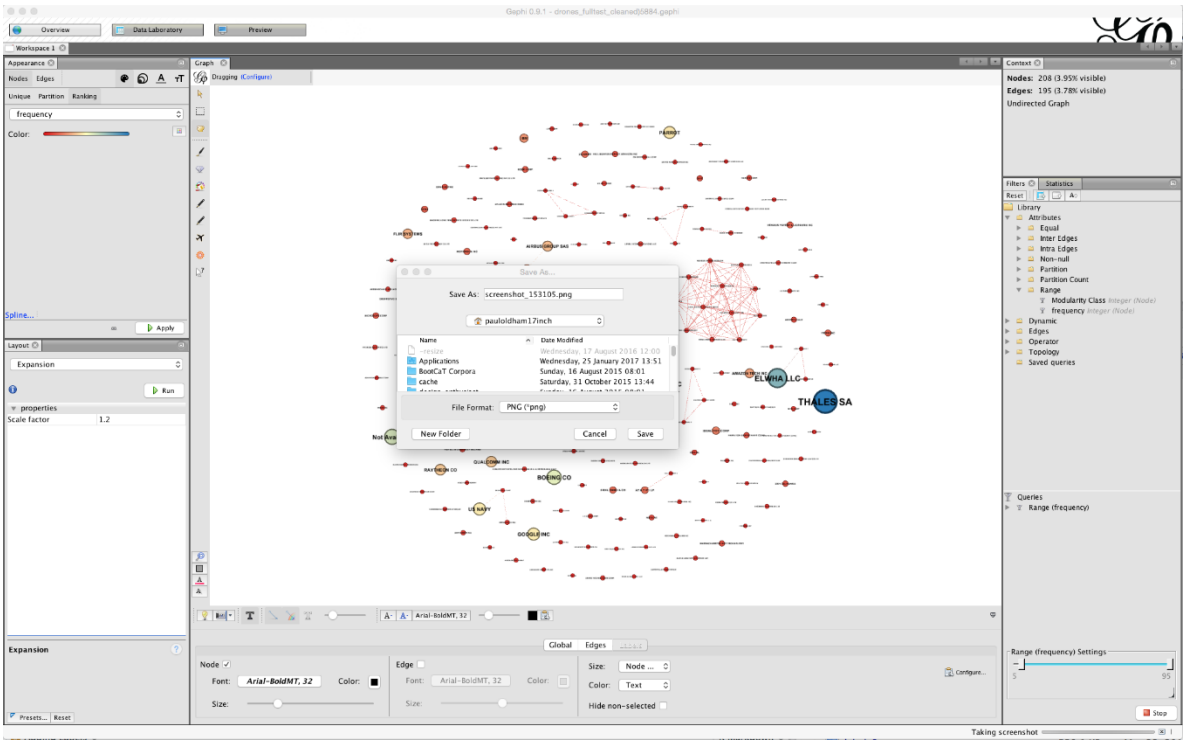
En la imagen de abajo, hemos usado Expansión dos veces y luego redimensionamos manualmente las etiquetas usando el control deslizante.

Ahora tendrás algo que se parece más o menos así.



Tenga en cuenta que puede utilizar el control deslizante a la derecha en el menú inferior para ajustar los tamaños y, por supuesto, podría ajustar la fuente. En algunos casos, puede estar contento con una red aproximada y lista en lugar de los ajustes detallados que se requieren para un gráfico de red final.

Tenga en cuenta el pequeño icono de la cámara a la izquierda del menú inferior. Oprima eso para tomar una captura de pantalla o mantenga presionado para abrir un menú de configuración que le permitirá elegir un tamaño.



Si opta por esta opción, es posible que también desee ajustar la fuente o el color y utilizar el menú inferior para obtener un resultado con el que esté satisfecho. En algunos casos (como trataremos más adelante), mover los nodos manualmente le permitirá llegar a una red más limpia para hacer una captura de pantalla.

Las capturas de pantalla pueden ser un paso muy útil para explorar datos o compartir datos internamente. Para obtener gráficos de calidad de publicación, deberá pasar a utilizar las Opciones de vista previa y participar en la gephi shufflelimpieza progresiva de la red para obtener un gráfico de calidad de publicación.

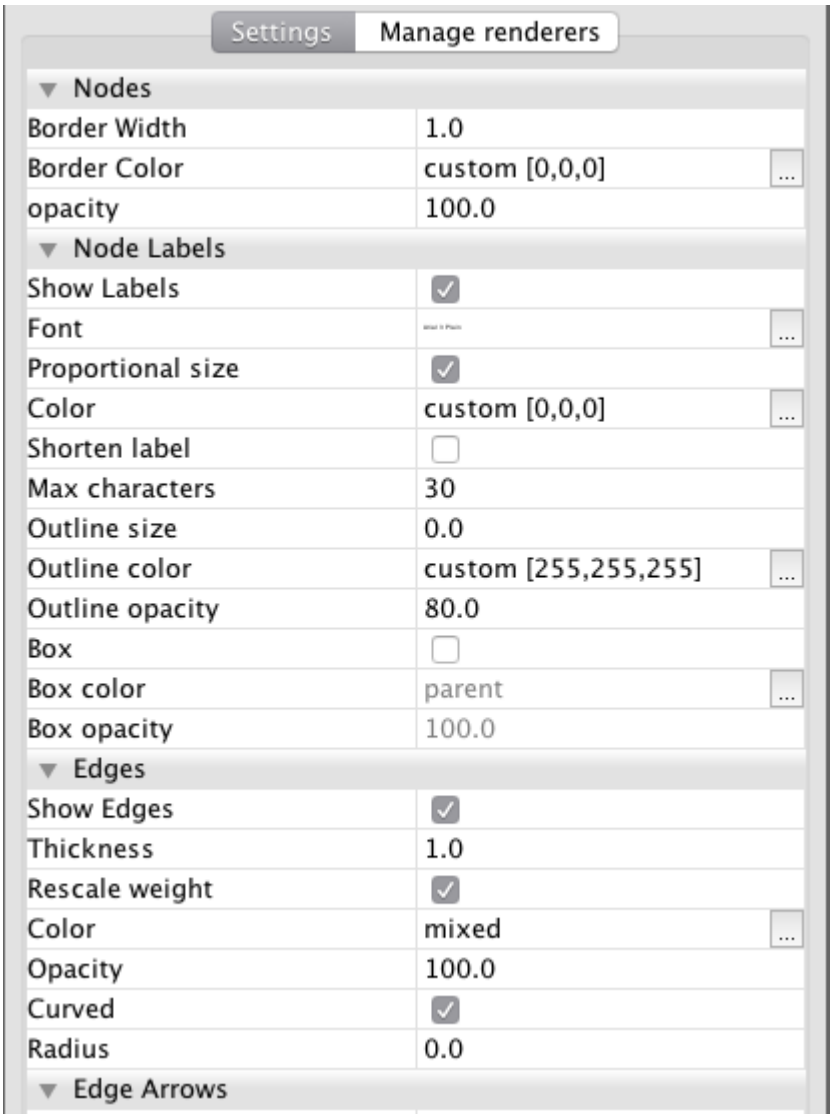
10.7 Usando las opciones de vista previa

Una opción más complicada para la visualización de la red es pasar a la pestaña Vista previa junto al Laboratorio de datos.

La opción predeterminada utiliza bordes curvos. Para utilizar esta prensa Refresh. Esto está bien, pero no podemos ver ninguna etiqueta. En los presets ahora intente por defecto curvo. Puedes jugar con las diferentes configuraciones hasta que encuentres una versión que te guste.

El principal problema que tenemos aquí es que las etiquetas son demasiado grandes y los pesos de línea también pueden ser muy pesados.

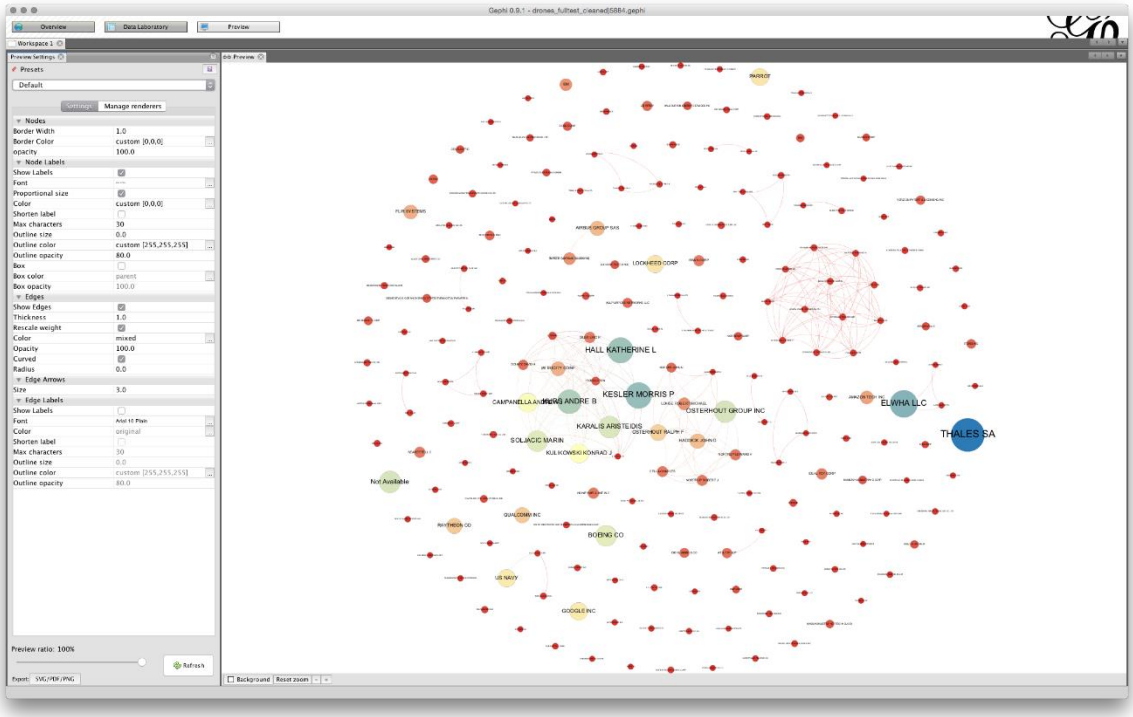
Para abordar el peso de la línea, busque y marque la opción de cambio de escala en los bordes.



Note aquí la diferencia con la visualización en el Resumen. Con Gephi lo que ves no es lo que obtienes.

Para llegar a una red más legible, la primera opción es ajustar el tamaño de la fuente en el Node Labelspanel de la configuración de vista previa. Tenga en cuenta que el tamaño de la etiqueta se establece para que sea proporcional a la fuente (desmarque eso y experimente si lo desea). Si nos atenemos al tamaño de fuente proporcional, comenzaremos más pequeños y nos moveremos hacia arriba. Por ejemplo, si ajustamos el tamaño de fuente a 3, el tamaño de fuente proporcional se reducirá. Al decidir sobre el tamaño de fuente, una consideración importante será

cuántos nodos desea que sean legibles para el lector. En este caso, establecer el tamaño de fuente en 3 y esto produce una red bastante legible.



Ese es un gráfico bastante aceptable para ver los nodos más grandes. Sin embargo, tenga en cuenta que las etiquetas de algunos de los nodos se superponen a algunos de los otros nodos. Esto puede producir un aspecto muy desordenado. Cuanto mayor sea el tamaño de la fuente base, más saturado se verá el gráfico y es probable que necesite más ajustes.

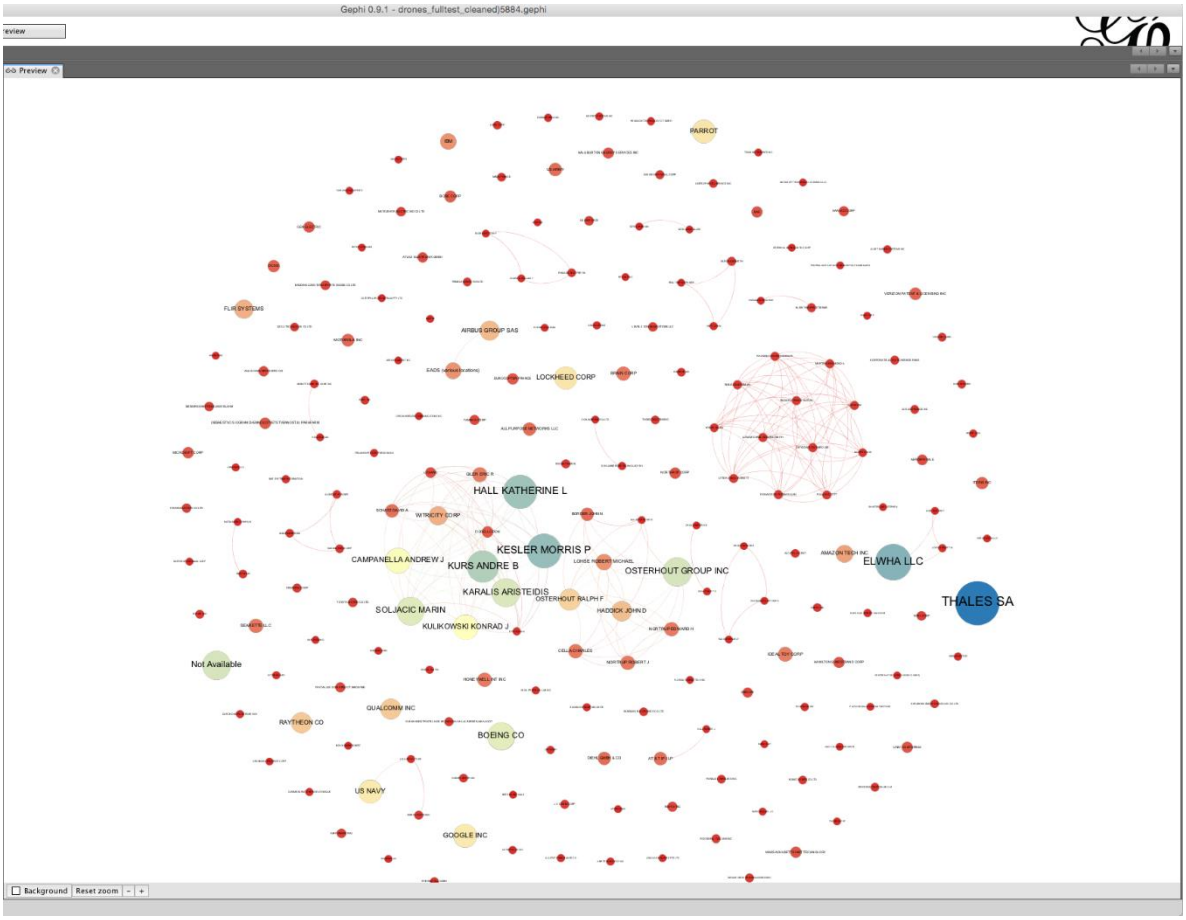
Para realizar ajustes en esta red, usaremos el tamaño 3. Ahora tendremos que avanzar y retroceder entre la Vista previa y la Descripción general ajustando la posición de los nodos. Para gráficos muy complejos, puede ayudar a imprimir la vista previa para ver qué necesita ajustar. Otra forma sensata de proceder es dividir mentalmente el gráfico en trimestres y avanzar en sentido horario trimestre por trimestre ajustando los nodos a medida que avanza. Es una muy buena idea guardar su trabajo en este punto y a medida que avanza.

En el primer y segundo trimestre, las cosas en movimiento en el sentido de las agujas del reloj se ven bien, sin etiquetas superpuestas. Sin embargo, se necesitan algunos ajustes en el tercer trimestre en la mitad de la red donde se superponen Campanella y Kurs. Para realizar el ajuste, muévase a la pestaña Información general, luego seleccione la manecilla pequeña en el menú vertical izquierdo para agarrar. Ubique Campanella y muévala fuera del camino para que no se

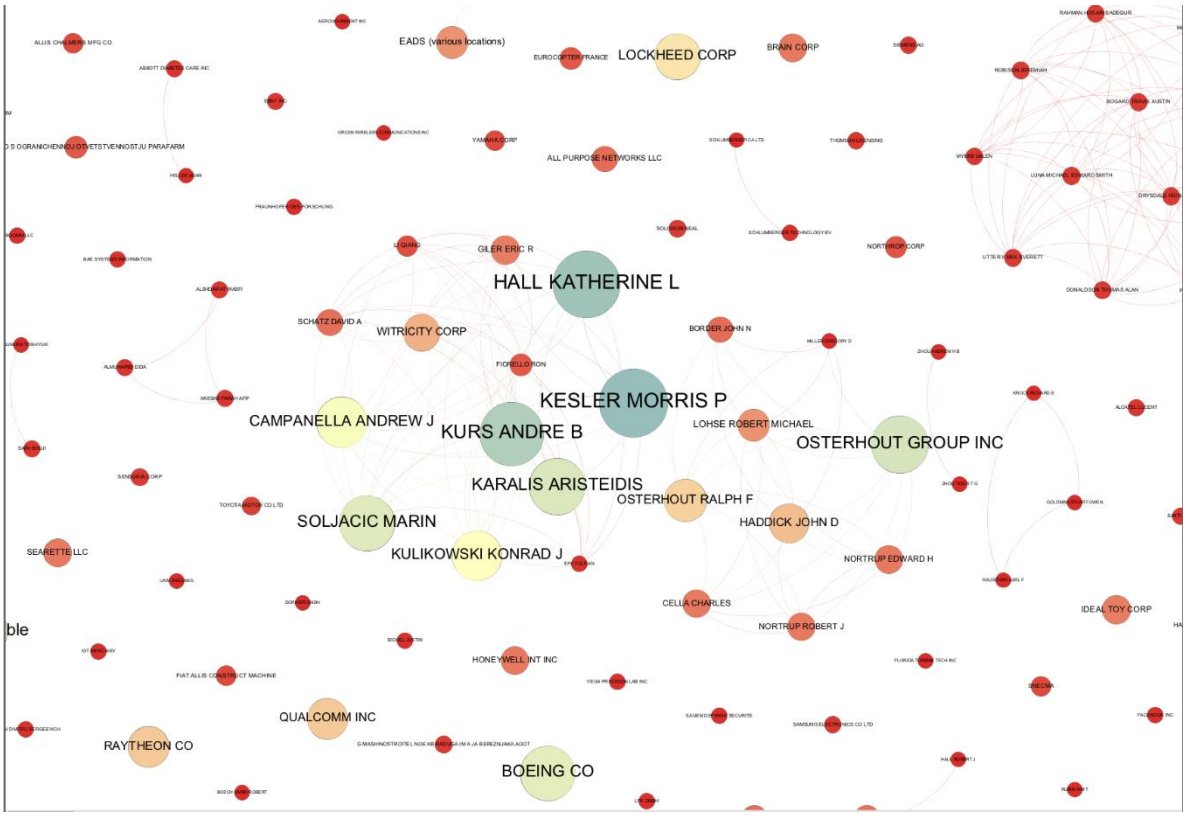
superponga. Sé gentil. Ahora vuelve a Vista previa y pulsa Actualizar. Al hacer este trimestre por trimestre, puede ser útil ampliar la vista general y la vista previa. Para cada uno de los nodos superpuestos trimestre a trimestre, realice un ajuste periódicamente para verificar nuevamente usando Actualizar en Vista previa y guardando a medida que avanza. Tenga en cuenta que el objetivo es realizar ajustes menores en lugar de ajustes importantes en la posición del nodo (también es posible intentar utilizar el ajuste de etiqueta en las opciones de diseño, pero en la práctica esto puede distorsionar la red). En el proceso, también vale la pena observar los bordes que se intersecan con los nodos donde no hay un enlace real. En esos casos, ajuste la posición del nodo intentando moverlo hacia el lado del borde no relacionado. Tenga en cuenta que a menudo esto no es posible con gráficos complejos y tendrá que explicar en el texto que los nodos pueden cruzarse con bordes no relacionados. También verifique que las ediciones de las etiquetas no contengan errores (como CATECH en lugar de CALTECH) y ajústelas según corresponda. Normalmente, las etiquetas largas causan problemas en este punto y se pueden editar en el Laboratorio de datos. En el proceso, también vale la pena observar los bordes que se intersecan con los nodos donde no hay un enlace real. En esos casos, ajuste la posición del nodo intentando moverlo hacia el lado del borde no relacionado. Tenga en cuenta que a menudo esto no es posible con gráficos complejos y tendrá que explicar en el texto que los nodos pueden cruzarse con bordes no relacionados. También verifique que las ediciones de las etiquetas no contengan errores (como CATECH en lugar de CALTECH) y ajústelas según corresponda. Normalmente, las etiquetas largas causan problemas en este punto y se pueden editar en el Laboratorio de datos. En el proceso, también vale la pena observar los bordes que se intersecan con los nodos donde no hay un enlace real. En esos casos, ajuste la posición del nodo intentando moverlo hacia el lado del borde no relacionado. Tenga en cuenta que a menudo esto no es posible con gráficos complejos y tendrá que explicar en el texto que los nodos pueden cruzarse con bordes no relacionados. También verifique que las ediciones de las etiquetas no contengan errores (como CATECH en lugar de CALTECH) y ajústelas según corresponda. Normalmente, las etiquetas largas causan problemas en este punto y se pueden editar en el Laboratorio de datos. Tenga en cuenta que a menudo esto no es posible con gráficos complejos y tendrá que explicar en el texto que los nodos pueden cruzarse con bordes no relacionados. También verifique que las ediciones de las etiquetas no contengan errores (como CATECH en lugar de CALTECH) y ajústelas según corresponda. Normalmente, las etiquetas largas causan problemas en este punto y se pueden editar en el Laboratorio de datos. Tenga en cuenta que a menudo esto no es posible con gráficos complejos y tendrá que explicar en el texto que los nodos pueden cruzarse con bordes no relacionados. También verifique que las ediciones de las etiquetas no contengan errores (como CATECH en lugar de

CALTECH) y ajústelas según corresponda. Normalmente, las etiquetas largas causan problemas en este punto y se pueden editar en el Laboratorio de datos.

A través de una serie de ajustes menores en el sentido de las agujas del reloj, debe llegar a un gráfico de red final. Espere pasar unos 20 minutos en la limpieza cuando esté familiarizado con Gephi, dependiendo del número de nodos. Vale la pena señalar que a menudo tendrá que volver para hacer los ajustes finales.



El principio básico aquí es que cada nodo debe tener una etiqueta legible cuando se acerca y que los bordes no deben intersecarse con los nodos no relacionados (excepto si esto es inevitable). En este caso, hemos tomado una captura de pantalla del núcleo de la red.



Es bastante común cuando se llega a un gráfico de calidad de publicación para descubrir repentinamente un error en la red. Por ejemplo, en la etapa de limpieza de datos es posible que haya decidido no agrupar dos compañías con nombres muy similares. Sin embargo, en la etapa de visualización de la red, la red sugiere que, en la práctica, las dos compañías son una y la misma. En este caso, verifique los datos y diríjase al Laboratorio de datos para agrupar los nodos en un solo nodo. Como esto sugiere, el tiempo dedicado a la limpieza y la preparación de los datos normalmente generará dividendos en términos de tiempo ahorrado más adelante en el proceso de análisis y visualización.

10.8 Exportando desde la vista previa

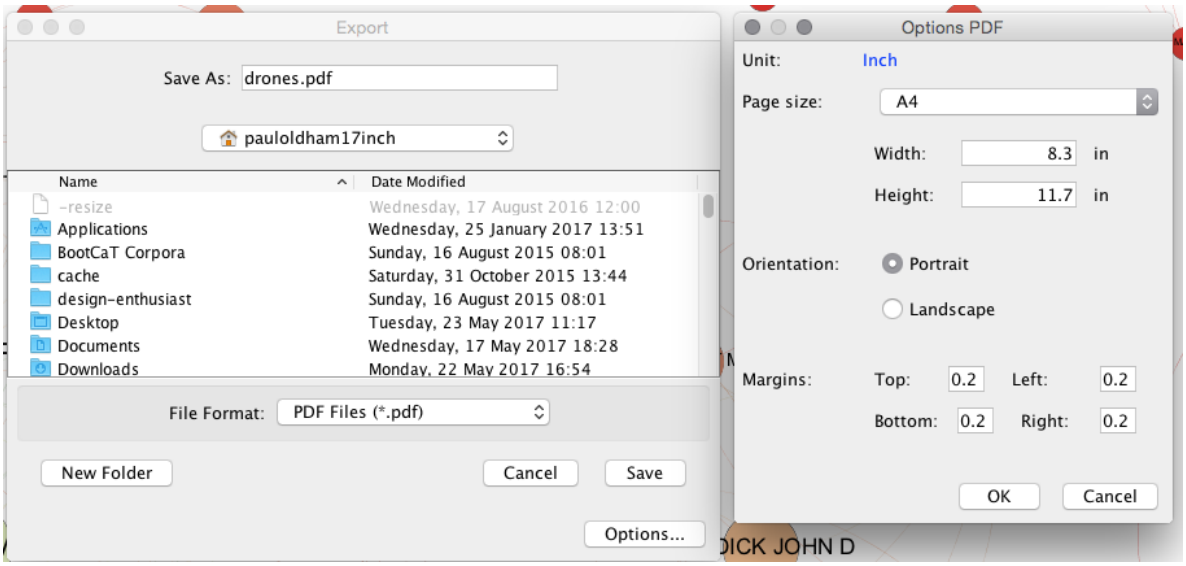
En esta etapa, desearemos hacer una verificación final y luego exportar los datos. Llegar a una exportación de calidad de publicación puede de hecho ser el aspecto más problemático y lento del uso de Gephi. Antes de seguir adelante guarda tu trabajo.

Al exportar, tenga en cuenta que lo que ve en la pantalla y lo que obtiene no es exactamente lo mismo. Los principales problemas son los bordes alrededor de los nodos y el peso de las líneas en los bordes. Para ajustarlo en el panel Nodos en Vista previa, cambie el ancho del borde a una configuración más baja o 0 (la opción

que seleccionamos aquí). En el panel de bordes, si no desea líneas gruesas, ajuste el grosor u opacidad (o ambos). En este caso, hemos reducido la opacidad de los bordes a 50 y hemos dejado el grosor como está. Si cambias algo, recuerda pulsar Refrescar.

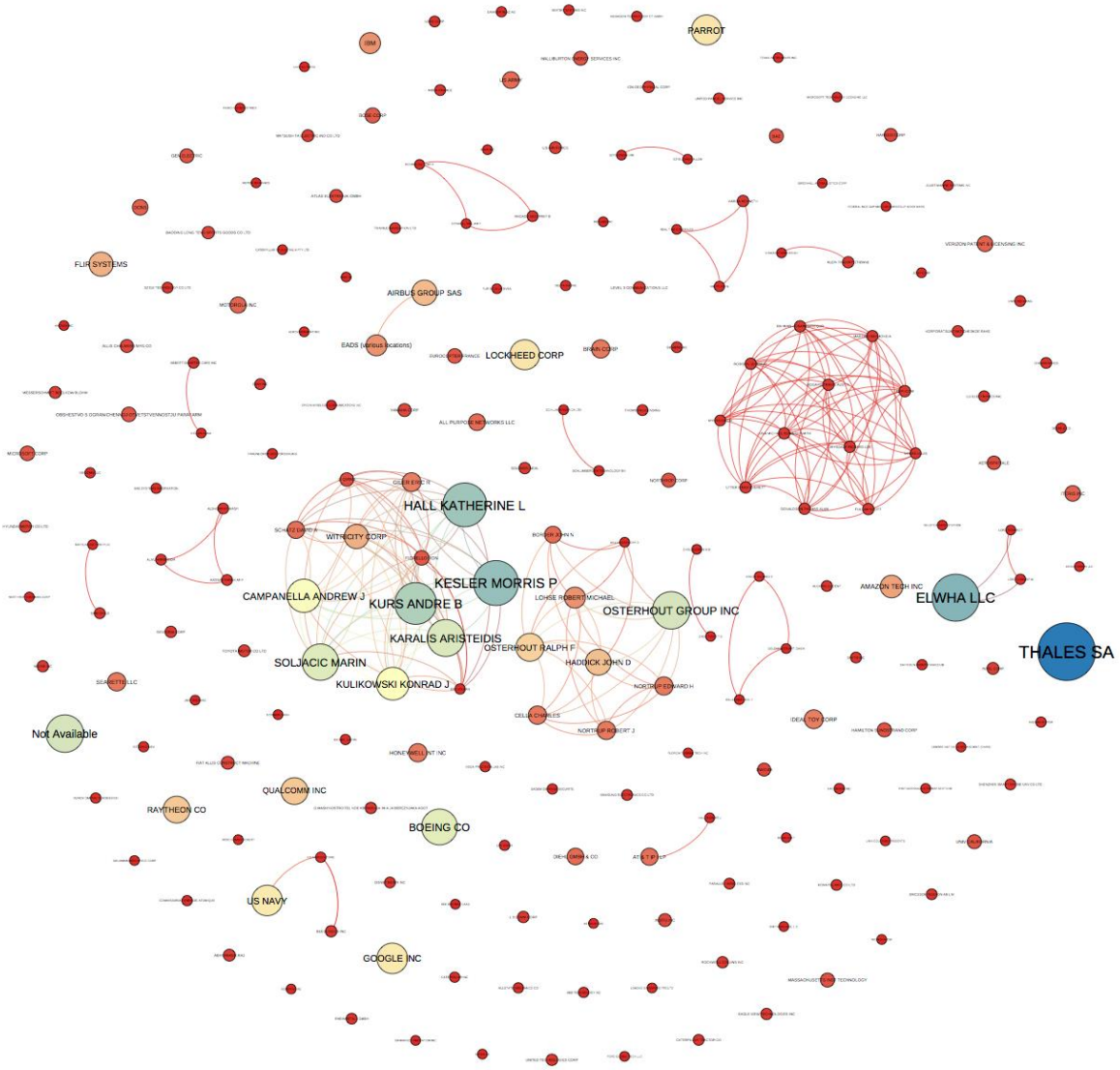
A continuación, seleccione el botón de exportación en la parte inferior izquierda. Exportaremos a .pdf.

Cuando elige Exportar, tenga en cuenta que hay un Optionsmenú para un control más estricto de la exportación.



Los valores predeterminados son razonables y los utilizaremos. Si está tentado de ajustarlos, tenga en cuenta que Gephi no recuerda sus configuraciones, incluso cuando se guardan, así que escríbalas. En realidad los valores por defecto funcionan muy bien.

Si todo va bien, terminarás con una imagen que se ve así.



Como el tamaño predeterminado es retrato, querrá recortar la imagen. Para la publicación, también querrá delinear el texto (para corregir la fuente en todo el sistema). Esto se puede hacer con el software gratuito GIMP o software de pago como Adobe Illustrator.

Enhorabuena, ahora has creado tu primer gráfico de red Gephi.

10.9 recursos

1. Sitio web de Gephi
2. Gephi repositorio github
3. De inicio rápido guía

4. Instrucciones de instalación para todas las plataformas . Gephi 8 sufre de un problema conocido para los usuarios de Mac. Es decir, utiliza Java 6, que no está instalado de forma predeterminada en Mac. Para resolver esto, debe seguir las instrucciones publicadas aquí y funciona muy bien en la mayoría de los casos. Básicamente, implica descargar una versión mac de Java que contiene Java 6 y luego ejecutar tres o cuatro comandos en la Terminal en el mac para configurar Gephi. Si eso no funciona, intente esta cuenta más detallada .
5. Convertidor de Excel / csv a plugin de red
6. Para obtener ideas sobre la visualización de la red de patentes, puede probar este artículo sobre biología sintética , este artículosobre los nombres de las especies en los datos de patentes y el uso del análisis exploratorio de redes utilizando el análisis de co-ocurrencia de IPC / CPC en el panorama de patentes de la OMPI para recursos genéticos animales. . Para más intente esta búsqueda de Google .

Capítulo 11 Patentes analíticas con Plotly

11.1 Introducción

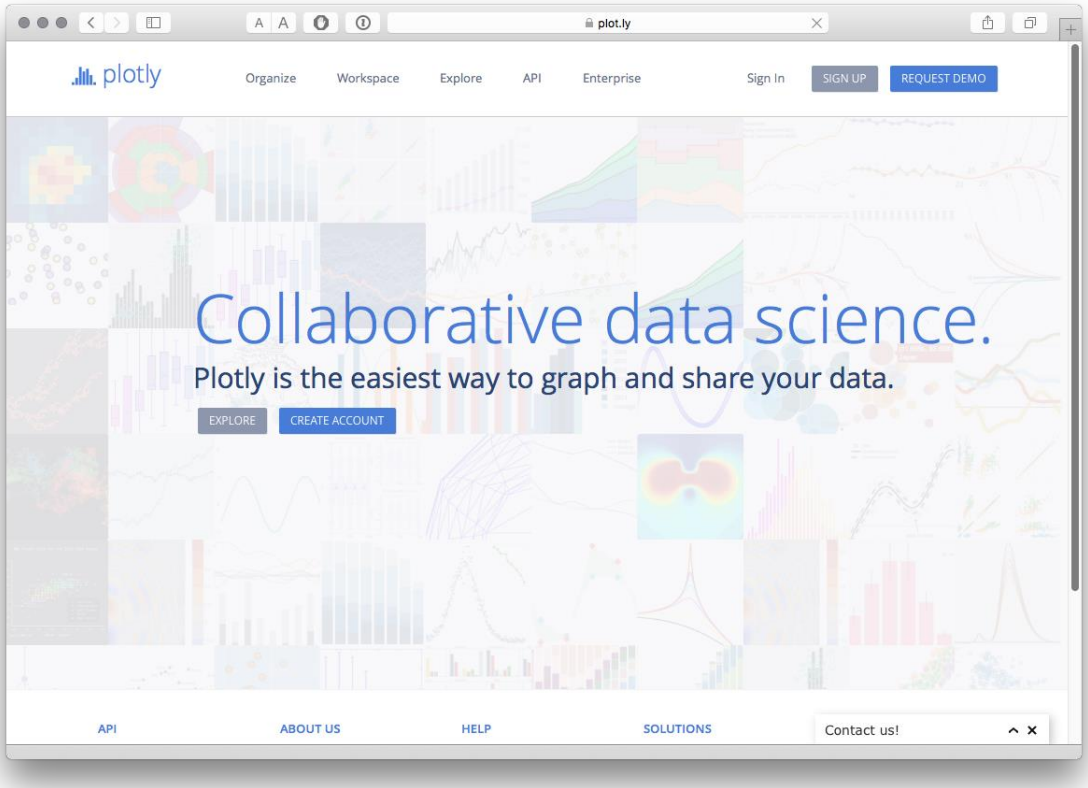
En este capítulo, proporcionamos una introducción al servicio de gráficos en línea Plotly para crear gráficos para su uso en el análisis de patentes.

Plotly es un servicio de gráficos en línea que le permite importar archivos de Excel, texto y otros para visualización. También tiene servicios API para R, Python, MATLAB y una biblioteca de Javascript Plotly. Una actualización reciente del plotly paquete en R le permite producir gráficos directamente en RStudio y enviarlos a Plotly en línea para su posterior edición y para compartir con otros.

La gran fortaleza de Plotly es que produce atractivos gráficos interactivos que se pueden compartir fácilmente con colegas o hacer públicos. También tiene una amplia variedad de tipos de gráficos, incluidos los mapas de contorno y de calor, y está construido con la popular biblioteca de Javascript D3.js para gráficos interactivos. Para ver ejemplos de gráficos creados con Plotly, consulte la galería pública. Plotly fue fundada en 2012 y, por lo tanto, es bastante nueva. Sin embargo, Plotly se está convirtiendo en una gran herramienta para crear y compartir gráficos. En este capítulo, nuestro objetivo es comenzar a utilizar Plotly en línea con archivos .csv o Excel. En la segunda parte del capítulo nos centraremos en utilizar el plotly paquete en RStudio para generar y exportar gráficos.

11.2 Primeros pasos con Plotly

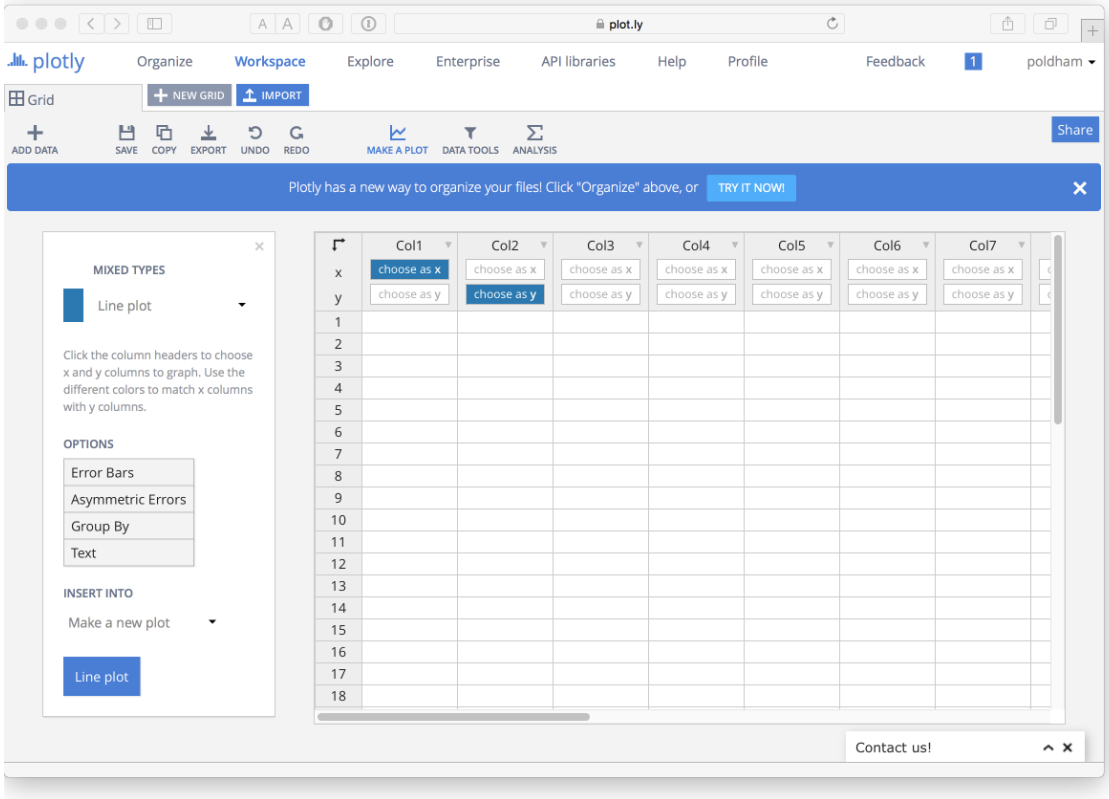
Necesitamos comenzar creando una cuenta usando el botón Crear cuenta.



A continuación, veremos una invitación para realizar una visita guiada (que vale la pena hacer) y Plotly señala que podemos cargar archivos de Google Drive o Dropbox. Luego seleccionamos la Workspaceopción para comenzar a trabajar.

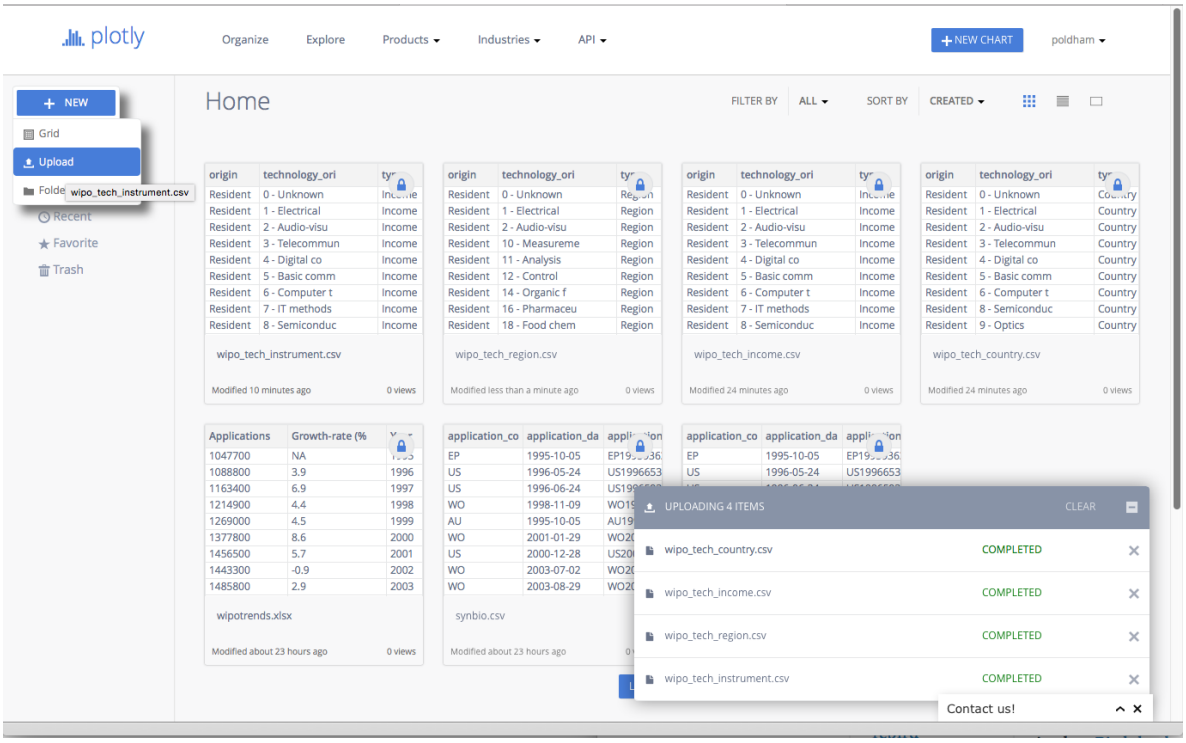
11.3 Importando archivos

Cuando llegue por primera vez, verá un área de trabajo con una cuadrícula (el término de Plotly para una tabla o hoja de trabajo).



En el espacio de trabajo, verá un Icono de importación que proporciona una gama de opciones para importar datos. ¡No importes nada todavía! También puede copiar datos de un archivo y pegarlos en la cuadrícula.

Al momento de escribir, el motivo para no usar estas opciones en este momento es que, si bien los datos pueden importarse bien por primera vez, en otros casos no lo harán. Al utilizar las opciones de esta página, no recibirá información si falla una importación. También tuvimos problemas para guardar los datos que se habían pegado en la hoja de trabajo (incluso cuando parecía que funcionaba). Para evitar la frustración potencial diríjase a Organize.



Desde la página Organizar, seleccione el botón Nuevo y luego Cargar. Ahora seleccione su archivo local. Cuando cargue el archivo, se mostrará un mensaje de estado y, si todo va bien, verá un mensaje completo. Si no, se mostrará un mensaje en rojo que le informa que ha habido un problema (no está claro cómo solucionar estos problemas).

Para este experimento, utilizamos dos conjuntos de datos del repositorio de datos del Manual de Open Source Patent Analytics . Cuando utilice el repositorio Github, haga clic en el archivo de interés hasta que vea un View Rawmensaje. Luego haga clic derecho para descargar el archivo de datos desde allí. Puede descargarlos para su propio uso directamente desde los siguientes enlaces.

1. La OMPI presenta tendencias de aplicación por año y con% de variación.
2. Patentes de pizza por país y año . Este es un conjunto de datos simple que contiene recuentos de documentos de patente que contienen la palabra pizza de WIPO Patentscope, desglosados por país y año.

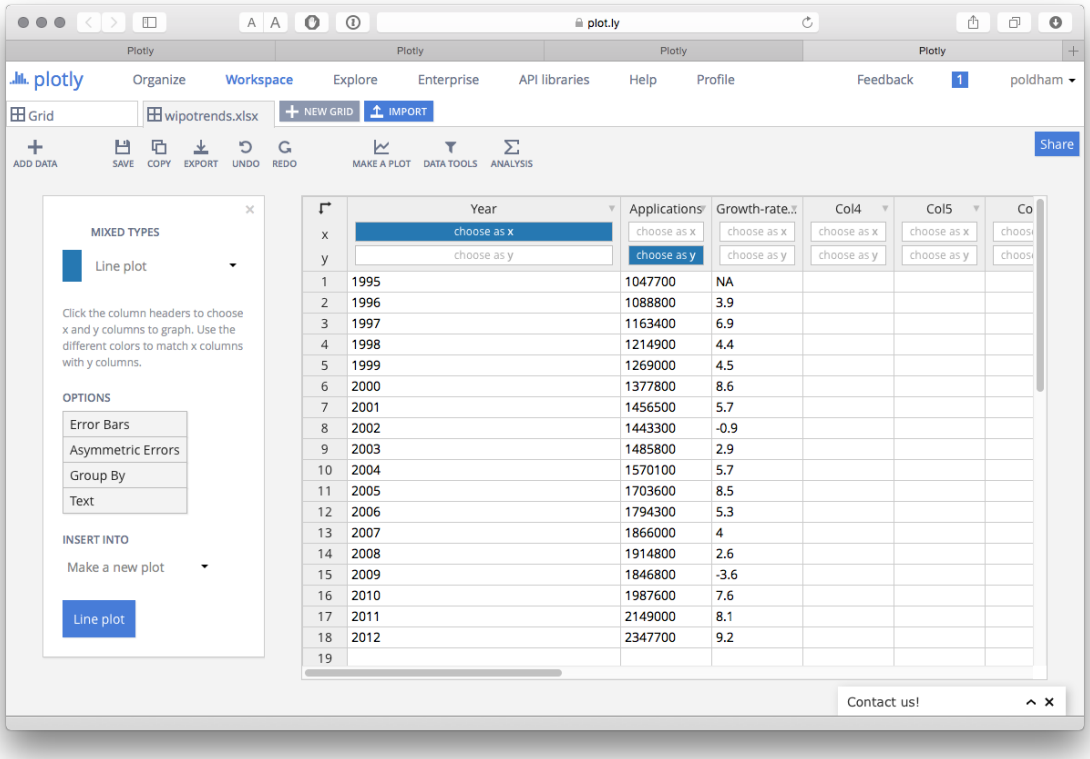
Un punto importante a tener en cuenta es que Plotly no es una herramienta de procesamiento de datos. Si bien hay algunas herramientas de datos, por lo general, sus datos deberán estar en una forma adecuada para trazar en el momento de la entrada. En parte, esto refleja el uso de API que permiten a los usuarios de Python, R y Matlab enviar sus datos directamente a Plotly para compartirlos con otros. Este es uno de los grandes puntos fuertes de Plotly y lo cubriremos a continuación. Sin

embargo, también tuvimos problemas al cargar y graficar conjuntos de datos con los que fue fácil trabajar en Tableau (como punto de referencia). Esto sugiere la necesidad de invertir tiempo en comprender los formatos que Plotly entiende.

Experimentamos un tipo diferente de problema con los simples datos de tendencias de la OMPI donde Plotly concatenó la primera fila (que contiene etiquetas) y la primera fila de datos en una fila de encabezado. Sin embargo, en la mayoría de los casos la importación parecía estar bien. Para convertir una fila en una fila de encabezado, haga clic con el botón derecho en la fila con los encabezados y haga clic derecho use row as col headers. Luego haz clic derecho nuevamente para eliminar la fila original.

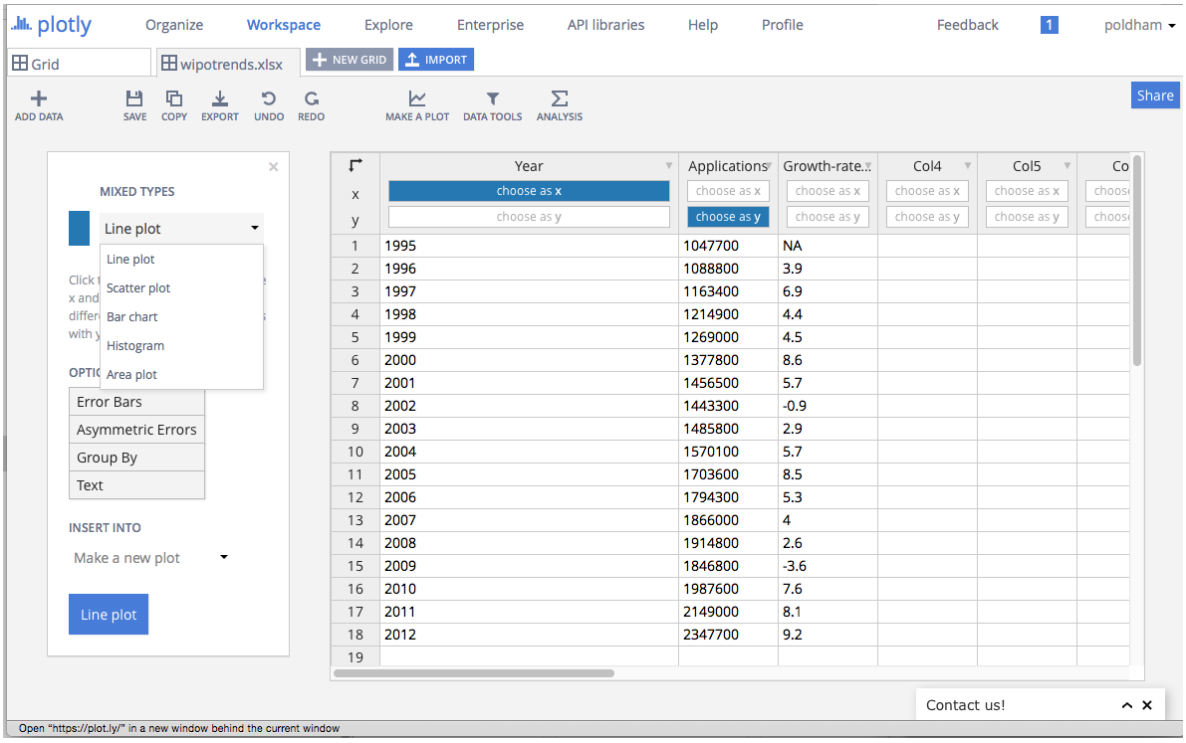
11.4 Creando un gráfico

Comenzaremos con los simples datos de tendencias de la OMPI abriendo esa Cuadrícula.

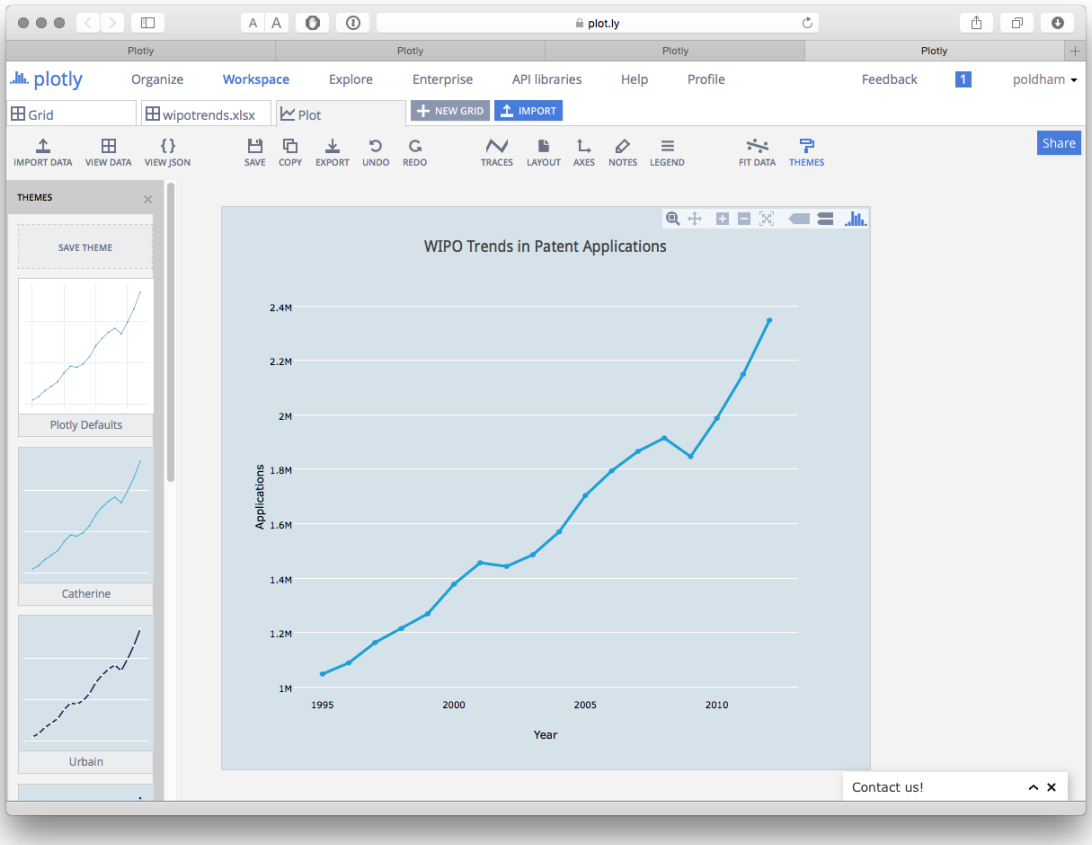


Tenga en cuenta que en la Cuadrícula tenemos opciones para seleccionar los ejes x o y para el trazado. También hay un menú de opciones al que volveremos.

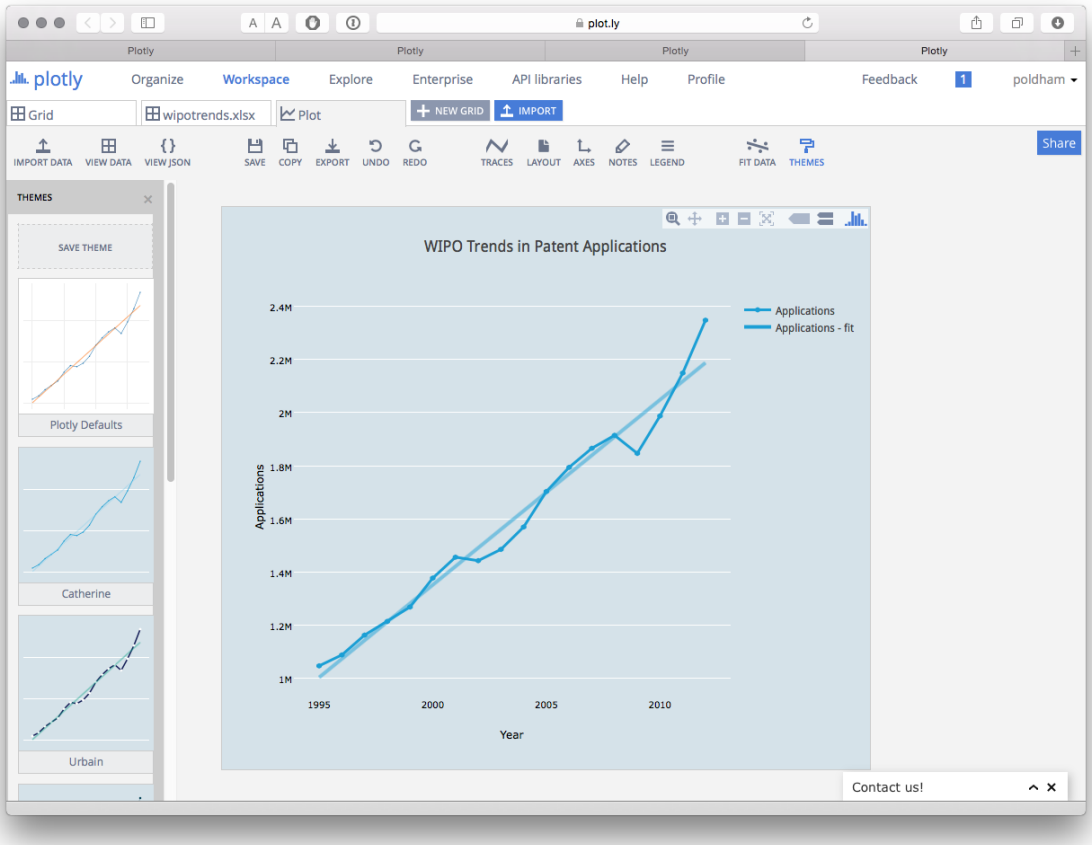
El tipo de gráfico se puede cambiar seleccionando el menú desplegable como podemos ver a continuación.



Siguiendo con un gráfico de líneas, cuando creamos la trama podemos agregar un título y luego cambiar el tema (en este caso a Catherine).



También podríamos agregar una línea de ajuste seleccionando el FIT DATAicono de menú. Esto le pedirá que cree un ajuste y luego tendrá un rango de funciones preestablecidas o puede agregar las suyas propias. Aquí simplemente hemos elegido el ajuste lineal predeterminado.

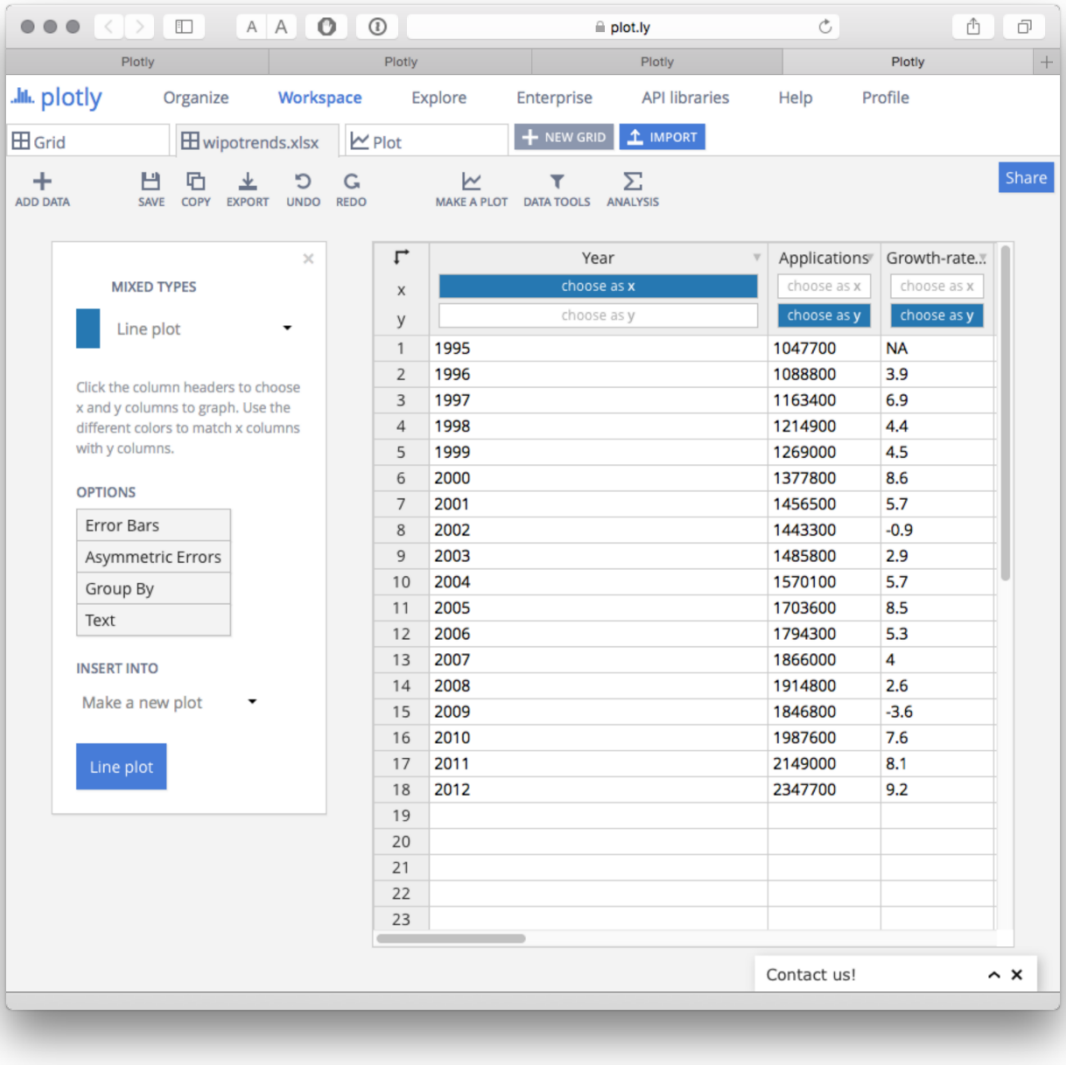


Luego podemos guardar la gráfica y usar el botón de exportación para guardar la gráfica en una variedad de formatos y tamaños. También es muy fácil agregar anotaciones usando el ícono de Notas. Confusamente, el gran botón azul Compartir solo parece guardar el archivo y, a pesar de guardar la trama, no pudimos ubicarlo de nuevo. Si bien Plotly ciertamente se ve bien y parece tener funciones atractivas, no es intuitivo y las dificultades que implica importar y compartir pueden ser frustrantes y llevar mucho tiempo. En resumen, se necesita tiempo para invertir y explorar el potencial de esta herramienta.

11.4.1 Añadiendo un segundo eje

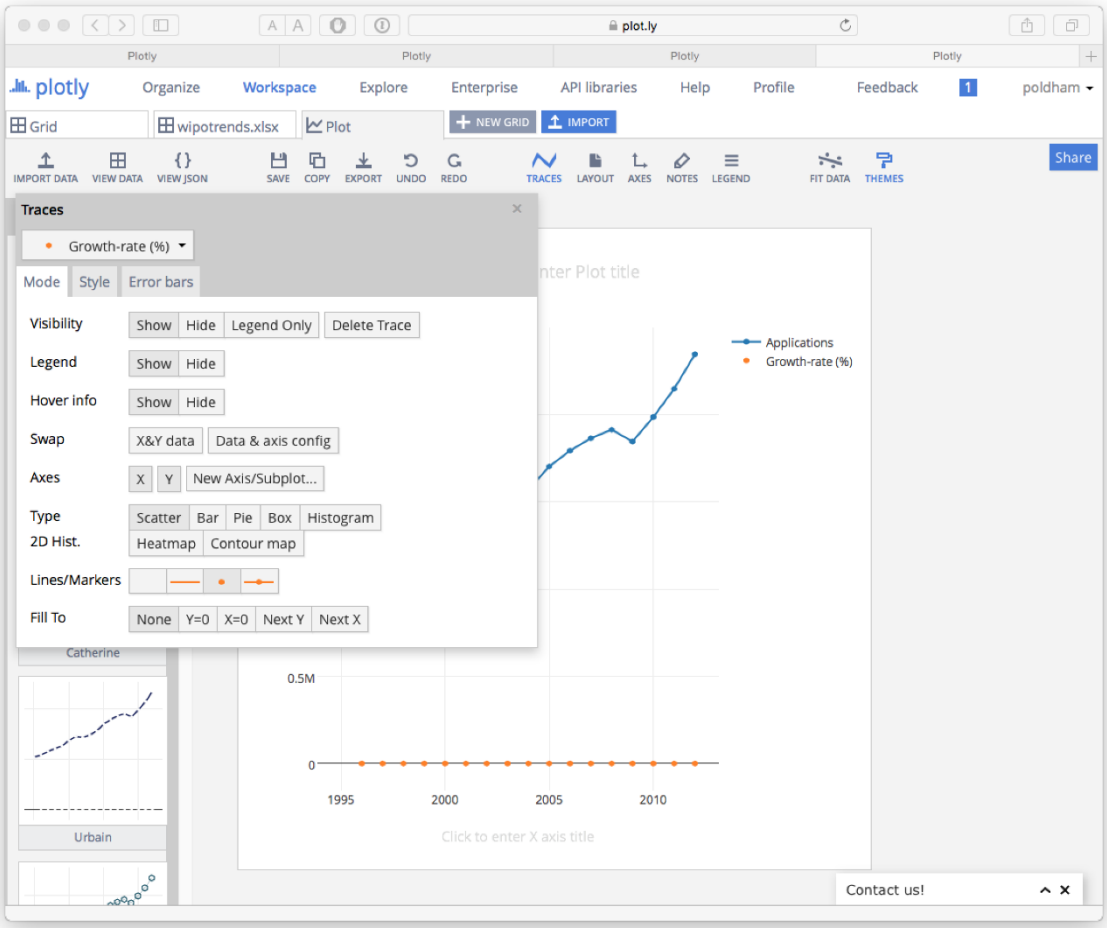
Si volvemos a nuestros datos originales de tendencias de la OMPI, tenemos un porcentaje de cambio anual en las solicitudes de patentes. Podríamos mostrar esto en un gráfico con un segundo eje para el porcentaje.

Para hacer eso en la vista de cuadrícula, seleccione el botón "elegir como y" en la columna de la tasa de crecimiento para agregar un segundo elemento para el eje y.

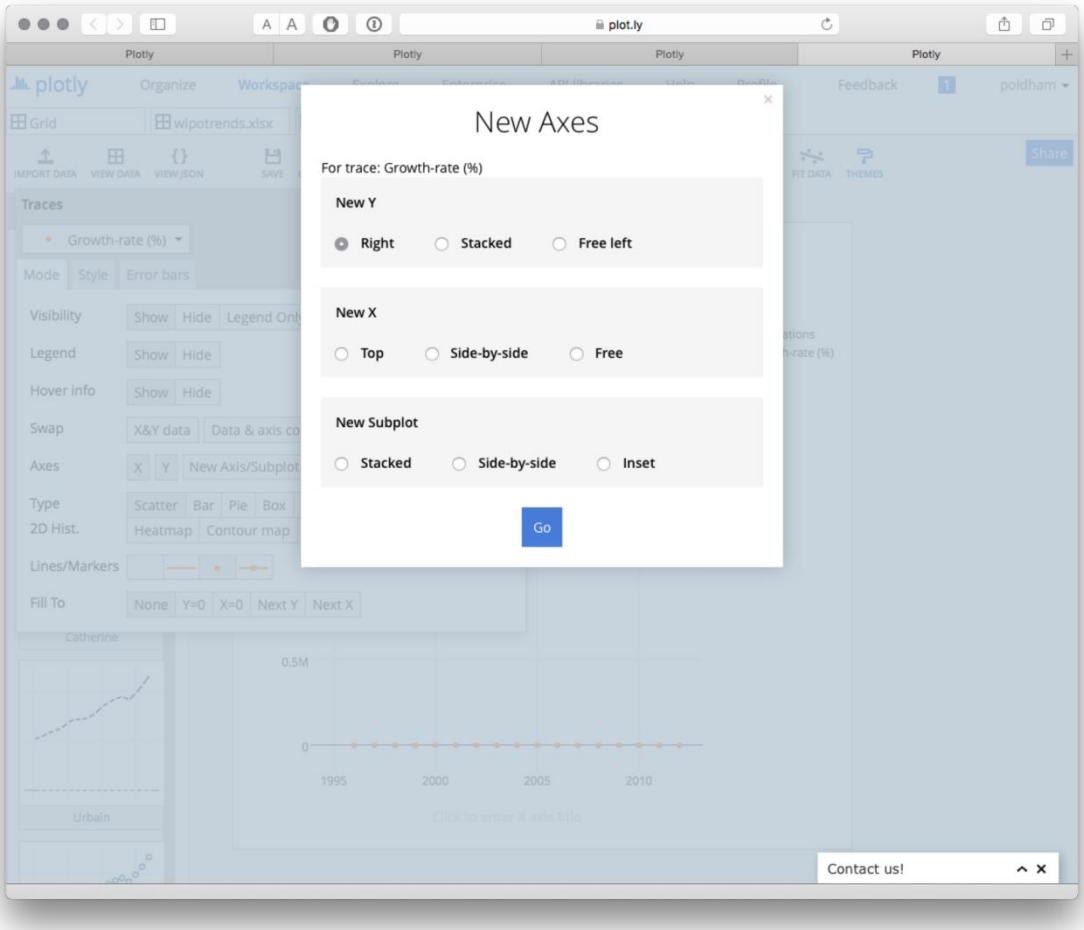


Cuando elijamos el gráfico de líneas, ahora veremos los dos conjuntos de datos con el porcentaje al final. Ahora necesitamos crear un segundo eje y a la derecha y asignar los datos de porcentaje a eso.

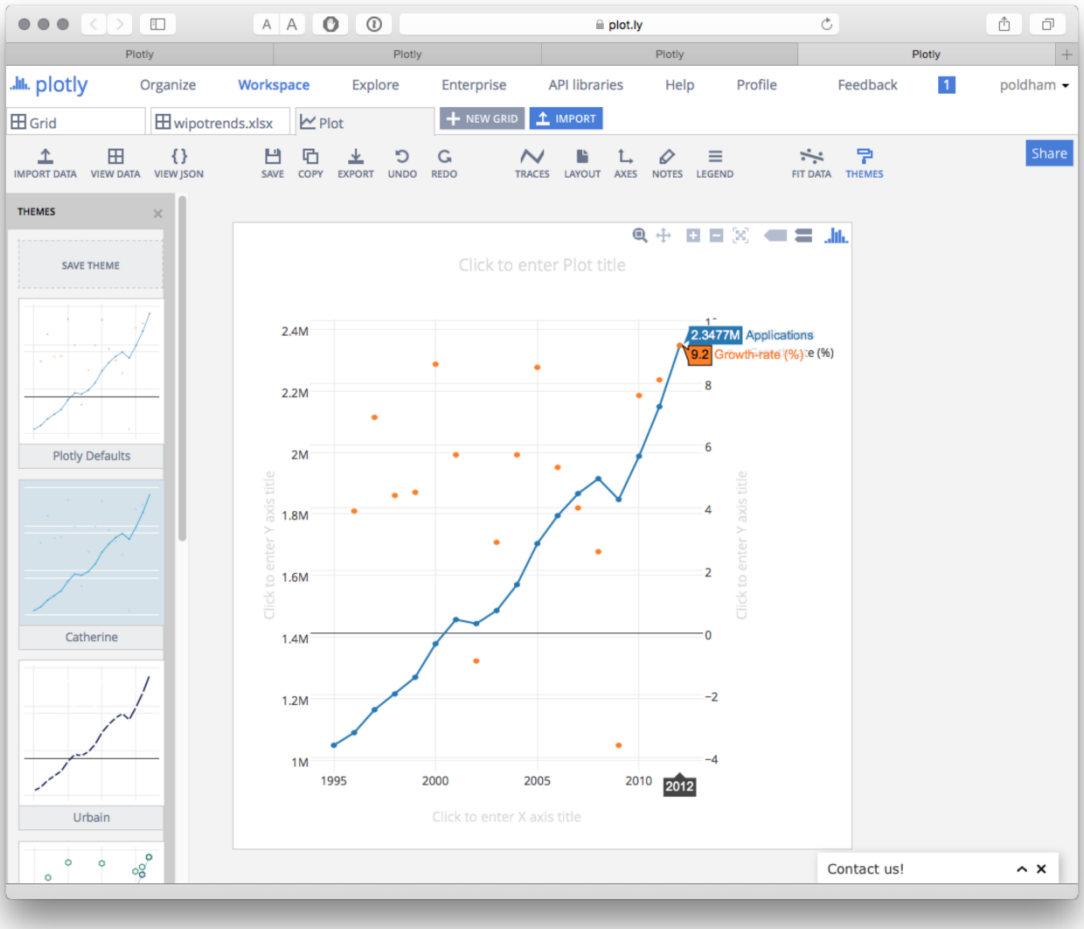
Para hacer esto, seleccione el icono de Trazas y aparecerá un menú que muestra las trazas de datos. El menú desplegable debajo de Trazas mostrará Aplicaciones, por lo tanto, seleccione% de tasa de crecimiento en el menú desplegable. Luego, donde veas Líneas / Marcadores, selecciona el punto. Esto evitará que las puntuaciones porcentuales se muestren como una línea.



A continuación, en el mismo panel bajo Axesseleccionar New Axis/Subploty aparecerá una nueva pantalla. Tenemos algunas opciones aquí, pero simplemente elegiremos crear un nuevo eje a la derecha.

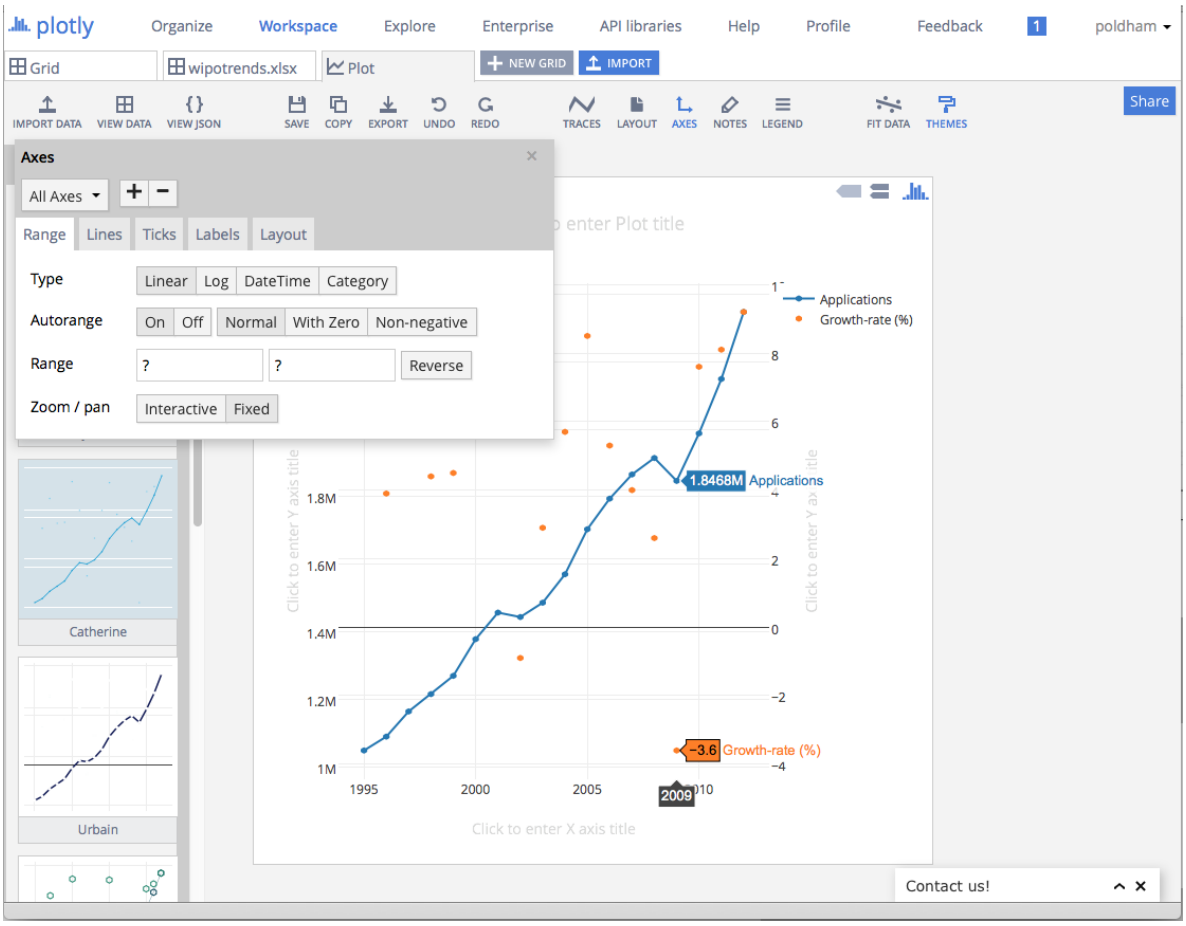


El resultado se verá algo como esto.



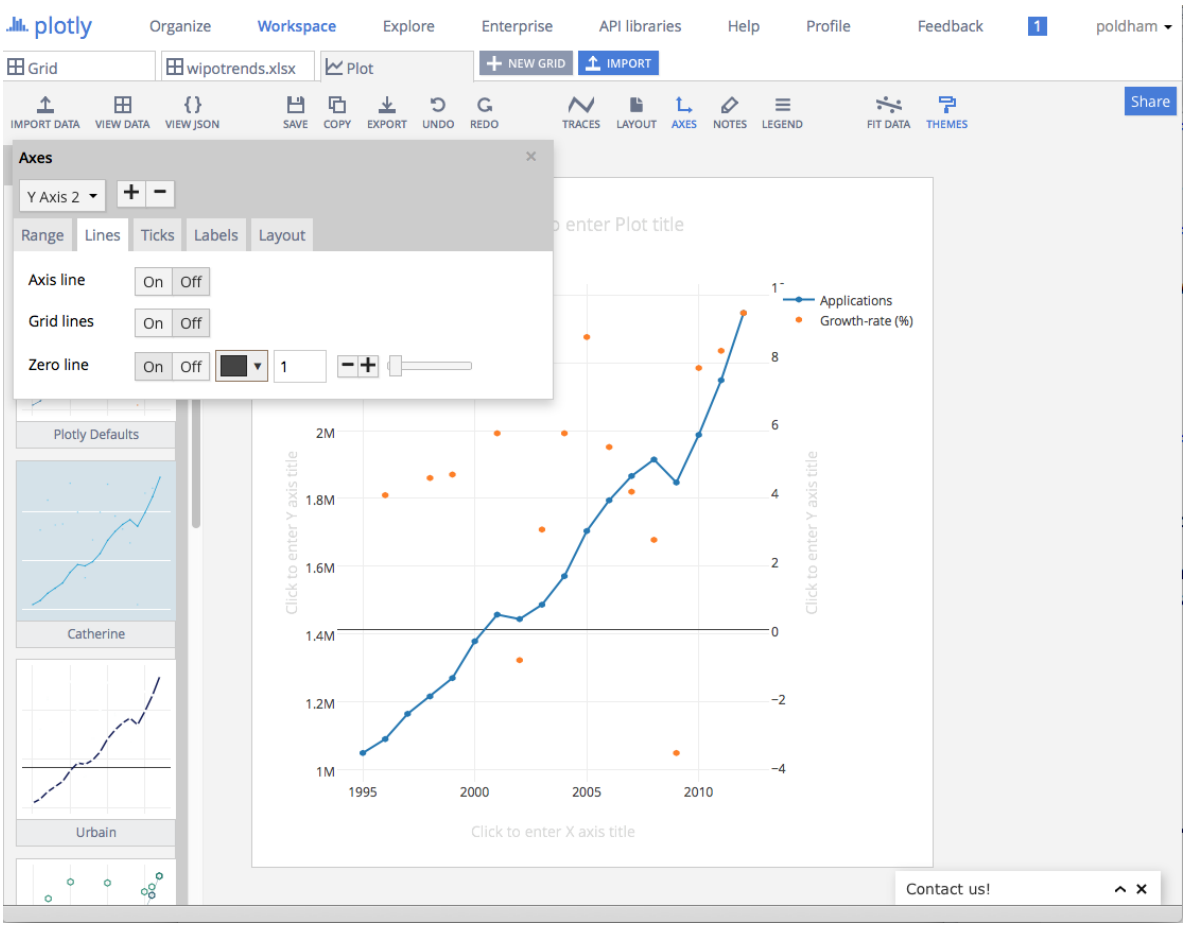
Nuestro problema ahora es igualar los ejes y cambiar el tamaño de los puntos para las puntuaciones porcentuales. Finalmente podemos añadir un título.

Antes de continuar, notemos que tenemos un valor de eje negativo significativo de -3.6% en 2009 cuando las solicitudes de patente declinaron. También hay un valor negativo en 2002.

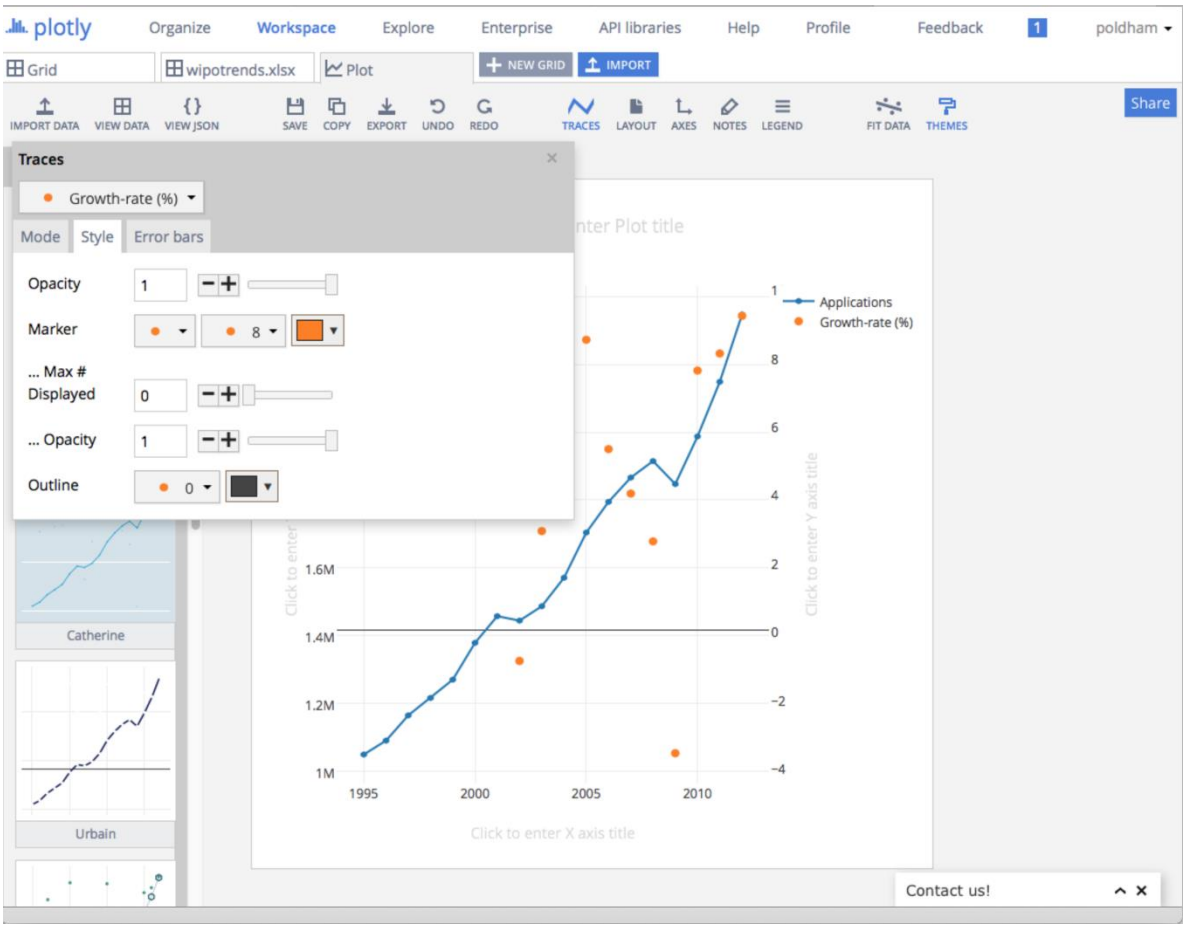


Si quisiéramos conservar estos valores, probablemente querríamos desactivar el segundo conjunto de líneas de cuadrícula. También nos gustaría cambiar el tamaño de los puntos.

Para desactivar las líneas de la cuadrícula en el segundo eje y, haga clic en el icono Ejes en el menú principal a la izquierda. Luego, en el All Axesmenú desplegable debajo de Ejes, seleccione Y Eje 2. Luego haga clic en el ícono del submenú Líneas y desactive las líneas y líneas de cuadrícula. También apague la línea Cero a menos que desee retenerla.



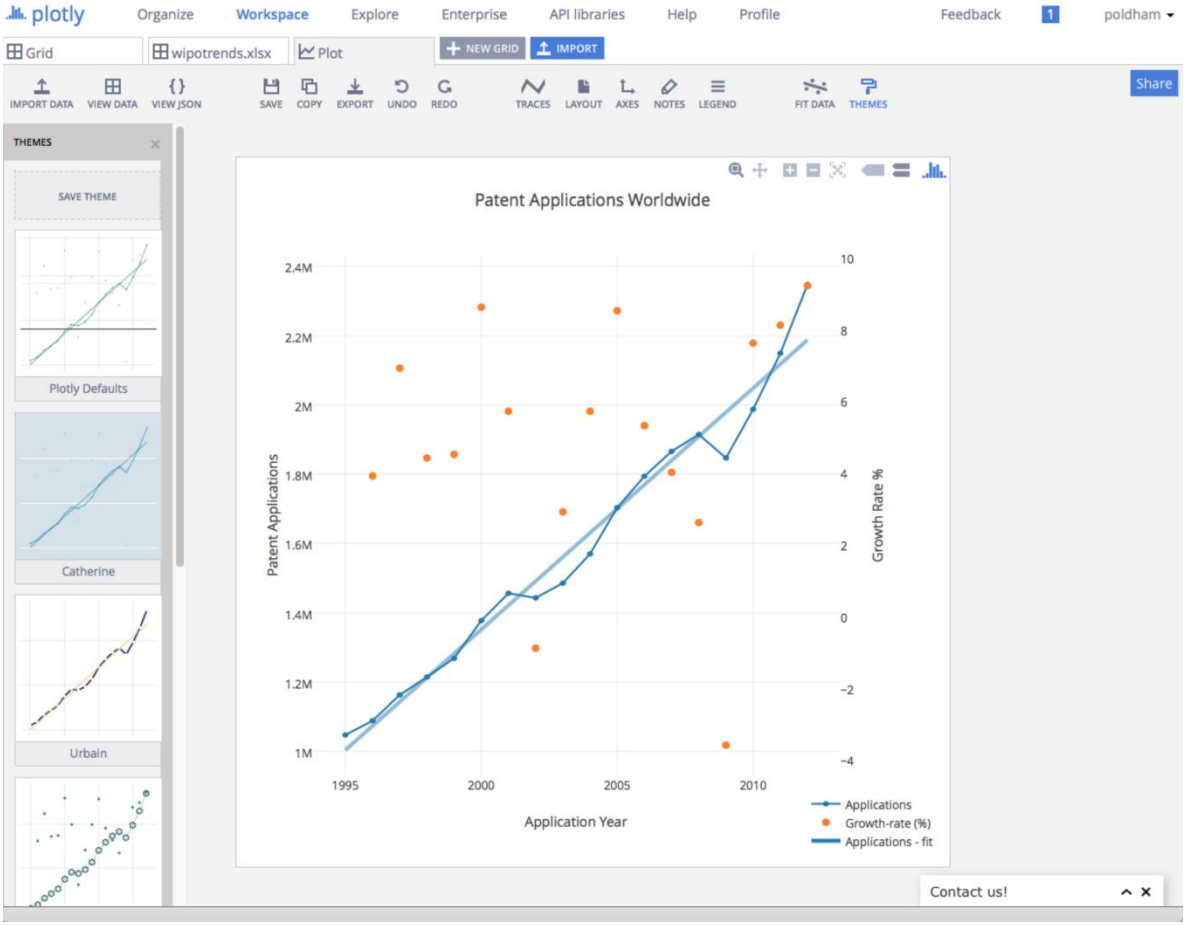
Para cambiar el tamaño de los puntos, debemos volver al menú principal de Trazas a la izquierda y seleccionar la Tasa de crecimiento de la lista de Trazas. Luego elija la pestaña Estilo y cambie el tamaño del marcador a algo más grande, como 8.



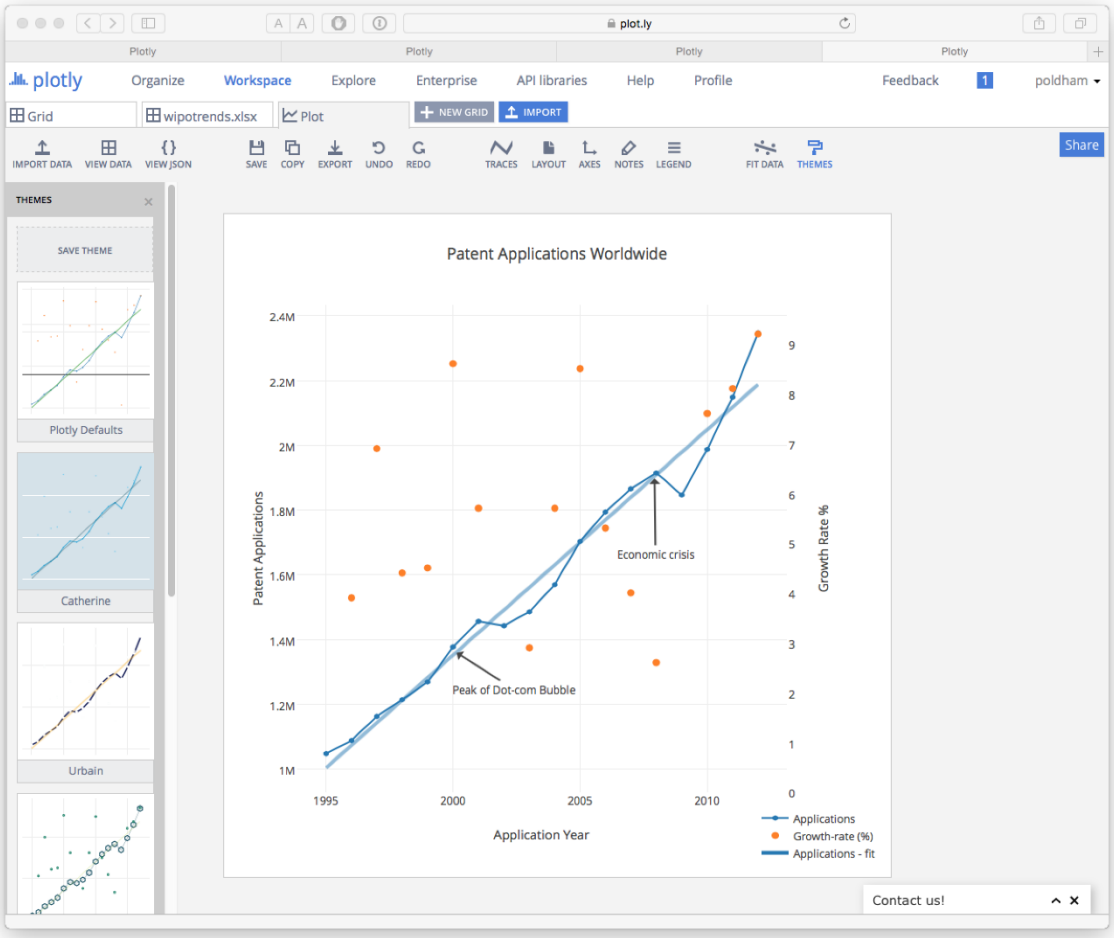
Para finalizar el gráfico queremos añadir algunas etiquetas. Simplemente podemos escribir las etiquetas del Eje y un título en los cuadros de texto provistos. Al elegir el ícono Leyenda, podríamos activar o desactivar la leyenda. Tenga en cuenta que, si bien este gráfico puede considerarse autoexplicativo, puede que no lo sea para el lector. También podemos simplemente arrastrar las etiquetas de los ejes a una posición diferente.

Es posible que quisiéramos eliminar los valores negativos de la gráfica (en ese caso, los valores deberán explicarse en el texto que lo acompaña). Para hacer eso Axes, seleccione y luego Y Axis2, en la Autorangeopción Non-negativede mostrar solo los valores mayores a cero en la gráfica.

Si quisiéramos, también podríamos aplicar una línea de ajuste seleccionando el icono Ajustar datos. Elegiremos lineales.



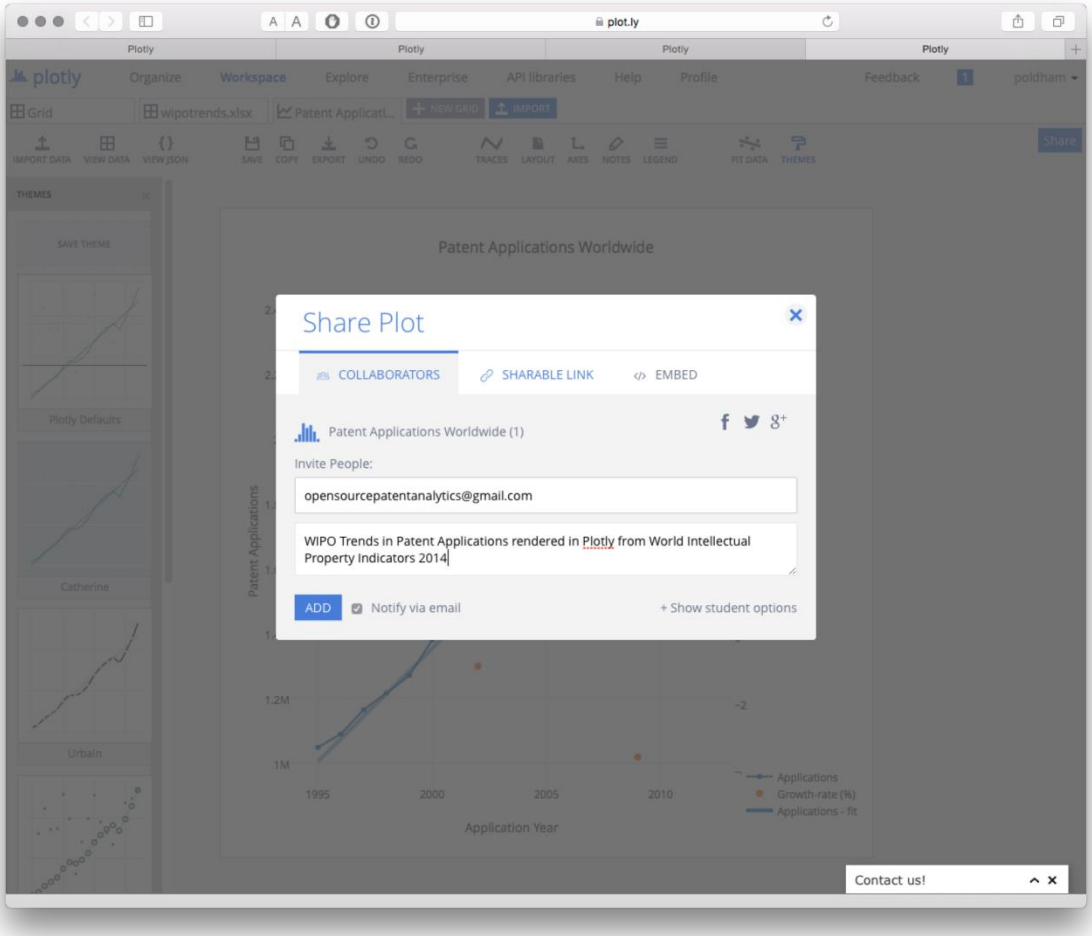
Finalmente, para terminar la trama, es posible que desee agregar anotaciones utilizando el icono NOTAS. Simplemente haga clic en el signo más en el menú emergente para obtener una nueva anotación y luego seleccione la flecha y el texto y muévalos a la posición que desee.



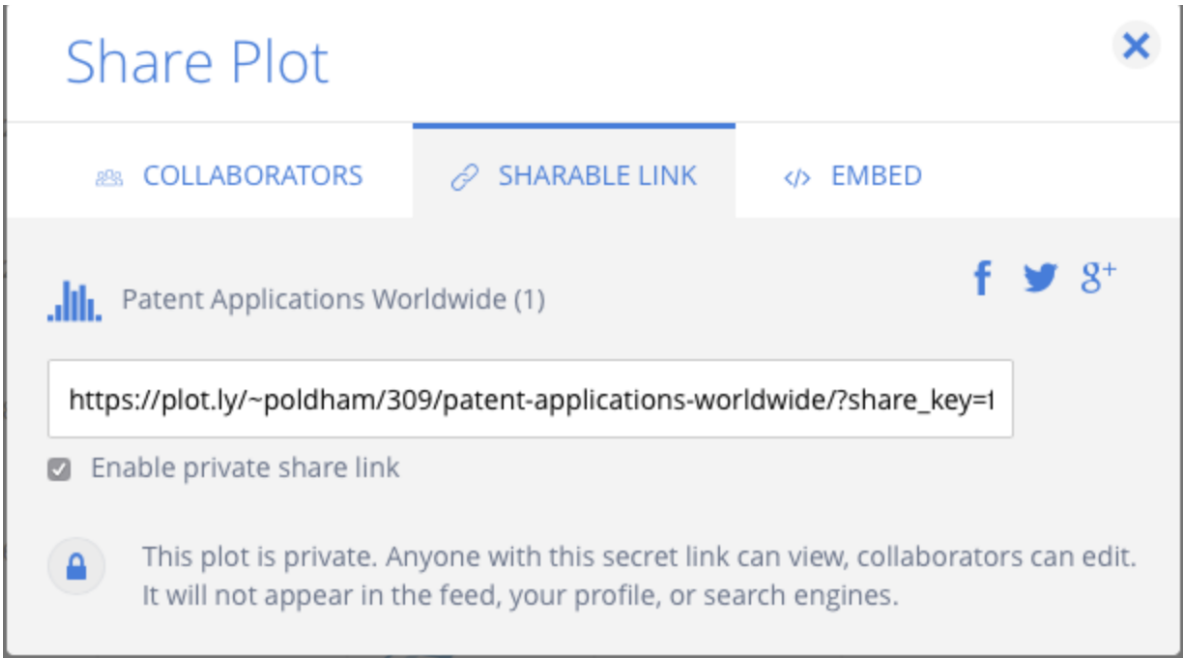
En este caso, hemos agregado un par de marcadores que pueden ayudar a comprender las tendencias en la actividad. Primero, tenemos una caída en las solicitudes de patentes entre 2001 y 2002. Una posible explicación aquí es que se trata de un golpe en el efecto del colapso de la burbuja de dot.com, donde los precios de las acciones alcanzaron un máximo en 2000, disminuyeron rápidamente y se recuperaron antes de disminuir. de nuevo en 2001. Los datos de patentes típicamente muestran efectos de retraso y es razonable pensar que la disminución en la actividad de la aplicación a partir de 2001 refleja estos ajustes económicos más amplios. Del mismo modo, hay una caída significativa en las solicitudes entre 2008 y 2009 que parece razonable asumir que refleja los efectos de la crisis económica mundial de 2007-2008. Tenga en cuenta que estos son marcadores de manera simple para ayudar a interpretar los gráficos.

11.5 Guardar y compartir

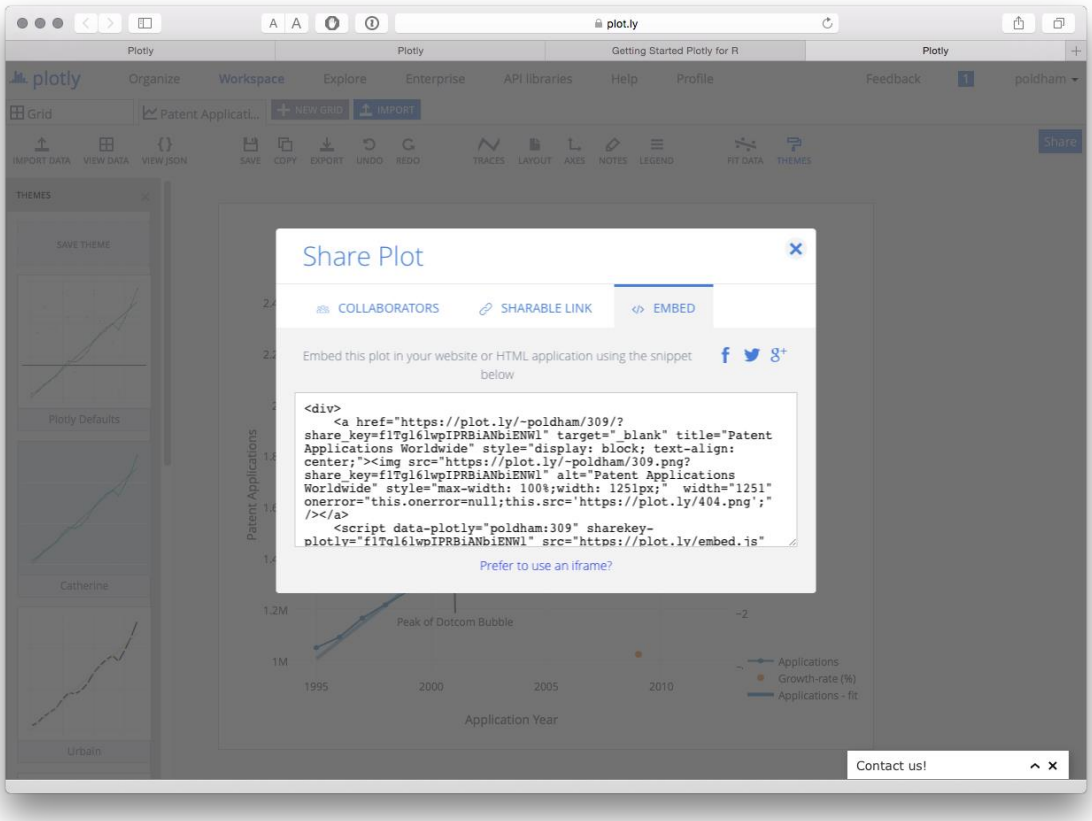
Para guardar la trama simplemente hacemos clic en Guardar. Sin embargo, es aquí donde una de las principales fortalezas de Plotly se hace evidente. Tan pronto como guardemos la trama también podemos invitar a otros por correo electrónico, podemos crear un enlace público o privado para compartir. Para los colaboradores, deben tener una cuenta de Plotly para que esto funcione.



La siguiente opción es compartir un enlace. Tenga en cuenta que el valor predeterminado es compartir un enlace privado. Para cambiar eso selecciona el icono de bloqueo. El enlace privado es particularmente adecuado para los profesionales de patentes.



También puede capturar un código de inserción para incrustar la trama en una página web



Alternativamente, sorprenda a sus amigos y familiares al publicar la trama en Facebook o comparta con una audiencia más amplia en Twitter.

En este ejemplo nos hemos centrado en desarrollar una trama muy simple utilizando plotly. En la práctica, hay una amplia gama de posibles opciones de trazado con un número creciente de tutoriales que se proporcionan aquí .

11.6 Trabajando con Plotly en R

Estamos siguiendo las instrucciones para configurar argumentalmente en I . Usaremos RStudio para este experimento. Descargue RStudio para su sistema operativo aquí y asegúrese de que también instala R al mismo tiempo desde el enlace en la página de RStudio aquí . Para Python, intente estas instrucciones de instalación para comenzar.

En RStudio primero tenemos que instalar el plotlypaquete. También instalaremos algunos otros paquetes útiles para trabajar con datos en R. Seleccione la pestaña Paquetes en RStudio e ingrese plotlye instale, o escriba lo siguiente en la consola y presione Entrar.

```
install.packages("plotly") # the main event
install.packages("readr")
 # import csv files
install.packages("dplyr")
 # wrangle data
install.packages("tidyr")
 # tidy data
```

A continuación, cargue las bibliotecas.

```
library(plotly)  library(readr)  library(dplyr)  library(tidyr)
```

Ahora necesitamos configurar nuestras credenciales para la API. Cuando inicie sesión, plotlysiga este enlace para obtener su clave API. Tenga en cuenta también que puede obtener un token de API de transmisión en la misma página. Streaming actualizará un gráfico desde el interior de RStudio.

Cuando haya obtenido su token, use el siguiente comando para almacenar su nombre de usuario y la clave API en su entorno.

```
Sys.setenv("plotly_username" = "your_plotly_username")
Sys.setenv("plotly_api_key" = "your_api_key")
```

A continuación, cargaremos un conjunto de datos de datos de la OMPI Patentscope que contienen datos de muestra en documentos de patente que contienen la palabra pizza organizada por país y año (pcy = pizza, país, año).

```
library(readr)
pcy <- read_csv
("https://github.com/wipo-analytics/
opensource-patent-analytics/raw/master/
2_datasets/pizza_medium_clean/pcy.csv")
```

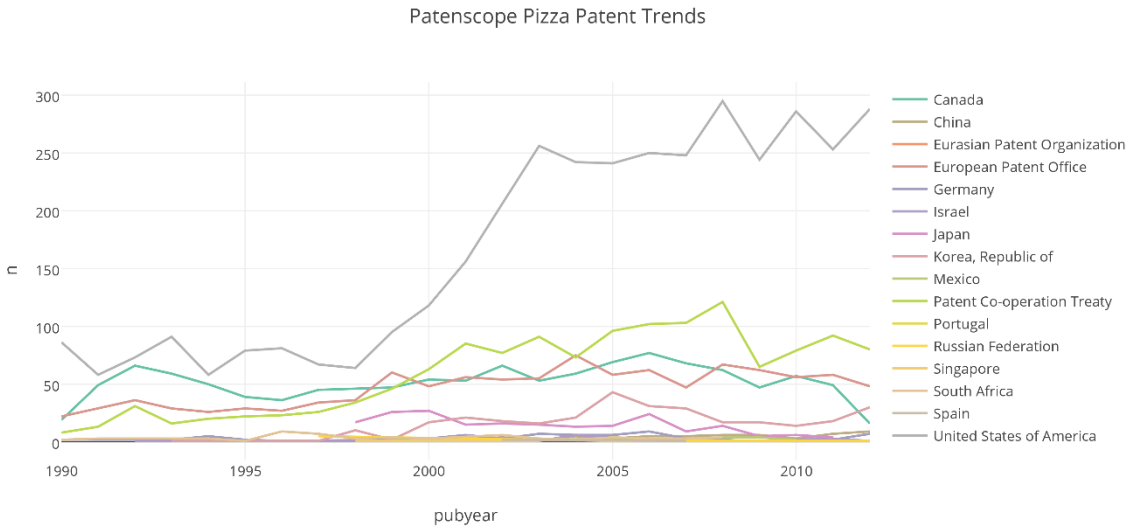
Debido a que los datos de patentes generalmente contienen un precipicio de datos para los últimos años, filtraremos los últimos años utilizando filter()del dplyrpaquete especificando un año menor o igual a 2012. Para eliminar la larga cola de datos históricos limitados, especificaremos mayor que o igual a 1990.

```
library(dplyr)
pcy <- filter(pcy, pubyear >= 1990, pubyear <= 2012) %>%
  print()

## # A tibble: 223 x 4
  ##    pubcountry pubcode pubyear     n
  ##    <chr>      <chr>     <int> <int>
##  1 Canada     CA         1990    19
  ##  2 Canada     CA          1991    49
  ##  3 Canada     CA          1992    66
  ##  4 Canada     CA          1993    59
  ##  5 Canada     CA          1994    50
  ##  6 Canada     CA          1995    39
  ##  7 Canada     CA          1996    36
  ##  8 Canada     CA          1997    45
  ##  9 Canada     CA          1998    46
  ## 10 Canada     CA          1999    47
  ## # ... with 213 more rows
```

Para crear la trama plotlyutilizamos la plot_ly()función. Especificaremos el conjunto de datos, los ejes x e y, luego, el color de los datos del país (conocido como traza en el plotlyidioma). Luego agregaremos un título usando el %>%operador de tubería para "esto" y luego "eso". Para especificar el aspecto

visual que queremos, especificamos el modo como "líneas" "(intente con" marcadores "para un diagrama de dispersión).

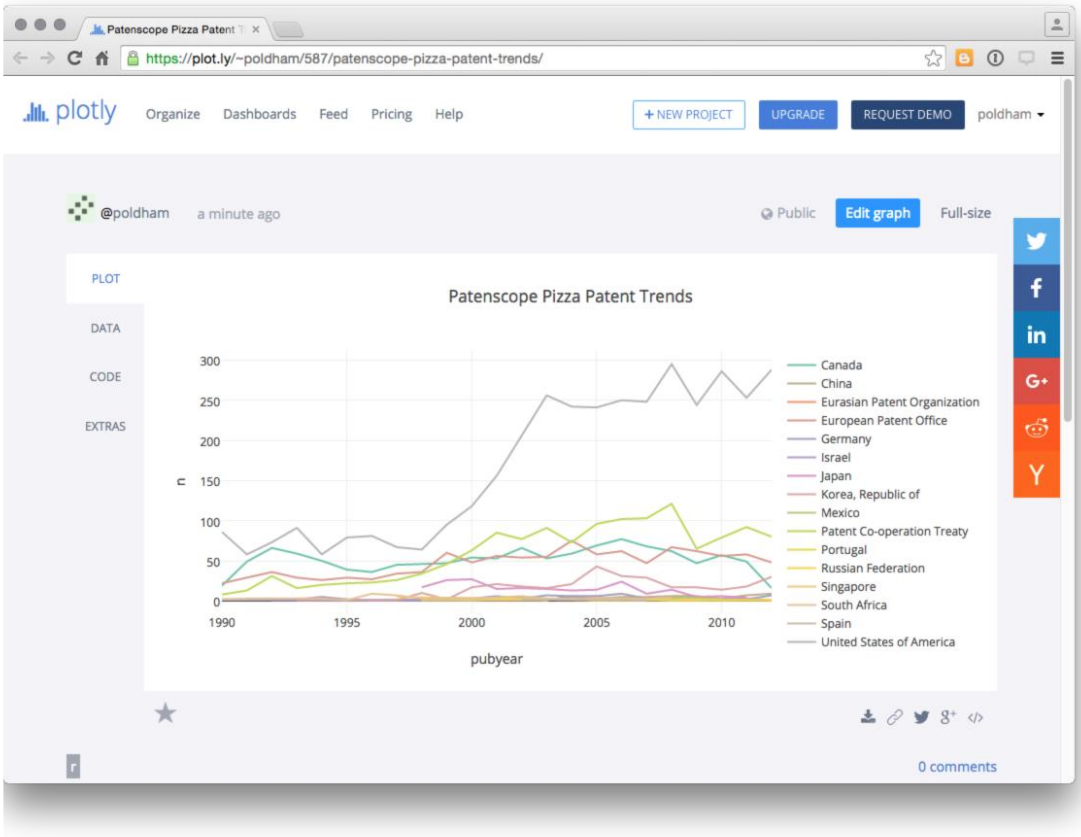
```
library(plotly)
s <- plot_ly(pcy, x = pubyear,
       y = n, color = pubcountry, mode = "lines") %>%
  layout(title = "Patenscope Pizza Patent Trends")s
```



Nuestros datos tienen más entradas que colores en la paleta de colores predeterminada. Plotly registrará una advertencia sobre el número de colores, pero ahora podemos ver claramente una trama. Si hemos ingresado nuestras credenciales para la API (arriba) también podemos insertar el gráfico en línea junto con los datos para su posterior edición o para compartir con otros.

```
library(plotly)  plotly_POST(s)
```

Esto abrirá una ventana del navegador y le pedirá que se registre o inicie sesión antes de acceder al gráfico.



Como queda claro, es fácil generar un plotlygráfico en R pero queremos profundizar en el plotlypaquete con un poco más de detalle.

Para cambiar los colores, es útil tener en cuenta que se plotlyinstala y luego llama al RColorBrewerpaquete (se mostrará en la lista de Paquetes). Para ver las paletas de colores, primero debemos marcar RColorBrewer en Paquetes (o library(RColorBrewer)) para cargarlo.

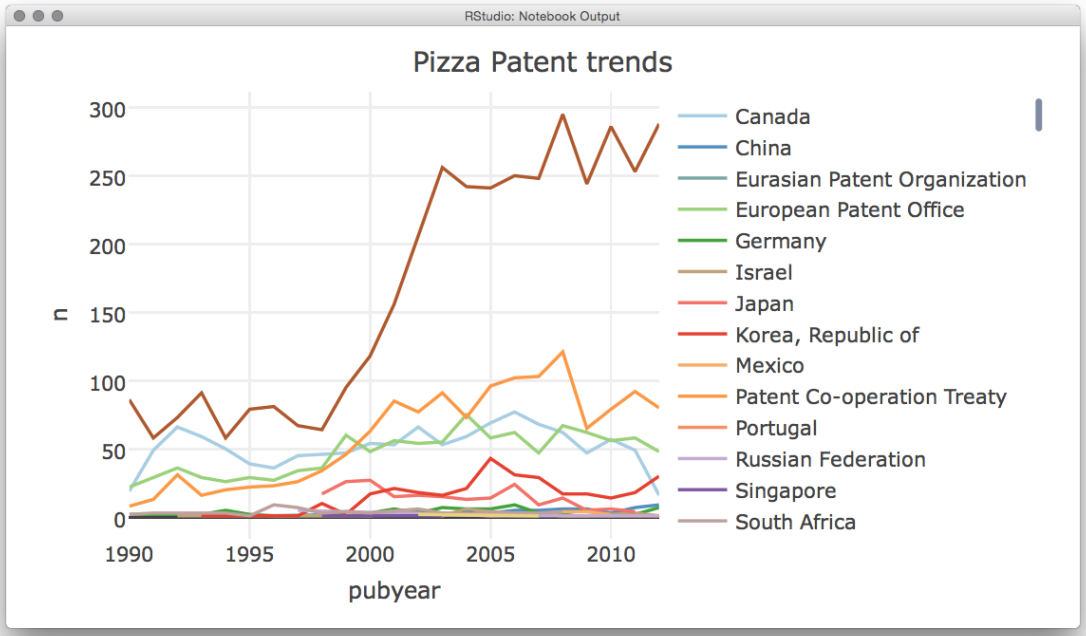
Para ver las paletas disponibles, simplemente puede utilizar View(brewer.pal.info)o la siguiente parte que organiza los datos por el número de colores.

```
library(RColorBrewer)
  library(dplyr)
  brewer.pal.info$names <- row.names(brewer.pal.info)
  select(brewer.pal.info, 4:1) %>%
    arrange(desc(maxcolors))
##        names colorblind category maxcolors
  ## 1     Paired       TRUE     qual        12
```

```
## 2       Set3      FALSE     qual         12
## 3       BrBG       TRUE      div         11
## 4       PiYG       TRUE      div         11
## 5       PRGn       TRUE      div         11
## 6       PuOr       TRUE      div         11
## 7       RdBu       TRUE      div         11
## 8       RdGy      FALSE      div         11
## 9     RdYlBu       TRUE      div         11
## 10    RdYlGn      FALSE      div         11
## 11  Spectral      FALSE      div         11
## 12   Pastel1      FALSE     qual          9
## 13      Set1      FALSE     qual          9
## 14     Blues       TRUE      seq          9
## 15      BuGn       TRUE      seq          9
## 16      BuPu       TRUE      seq          9
## 17      GnBu       TRUE      seq          9
## 18    Greens       TRUE      seq          9
## 19     Greys       TRUE      seq          9
## 20   Oranges       TRUE      seq          9
## 21      OrRd       TRUE      seq          9
## 22      PuBu       TRUE      seq          9
## 23    PuBuGn       TRUE      seq          9
## 24      PuRd       TRUE      seq          9
## 25   Purples       TRUE      seq          9
## 26      RdPu       TRUE      seq          9
## 27      Reds       TRUE      seq          9
## 28      YlGn       TRUE      seq          9
## 29    YlGnBu       TRUE      seq          9
## 30    YlOrBr       TRUE      seq          9
## 31    YlOrRd       TRUE      seq          9
## 32    Accent      FALSE     qual          8
## 33     Dark2       TRUE     qual          8
## 34   Pastel2      FALSE     qual          8
## 35      Set2       TRUE     qual          8
```

Esto indica que el número máximo de colores en una paleta es 12. Intentemos con Pairedfines ilustrativos. Esto tiene la ventaja de ser un color ciego amigable.

```
library(plotly)
  library(dplyr)
  s1 <- plot_ly(pcy, x = pubyear,
   y = n,color = pubcountry, colors = "Paired", mode = "lines") %>%
    layout(title = "Pizza Patent trends")  s1
```



Como podemos ver, esto producirá una gráfica con la paleta de colores, plotlymostrará una advertencia de que la paleta base ("Set2") tiene 8 colores, pero luego especificará que está mostrando la paleta que solicitamos.

En la práctica, nos gustaría dividir esta trama en subparcelas por dos razones. Primero, los rangos de datos y valores varían ampliamente entre países y segundo, es mejor asegurarse de que los colores sean distintos.

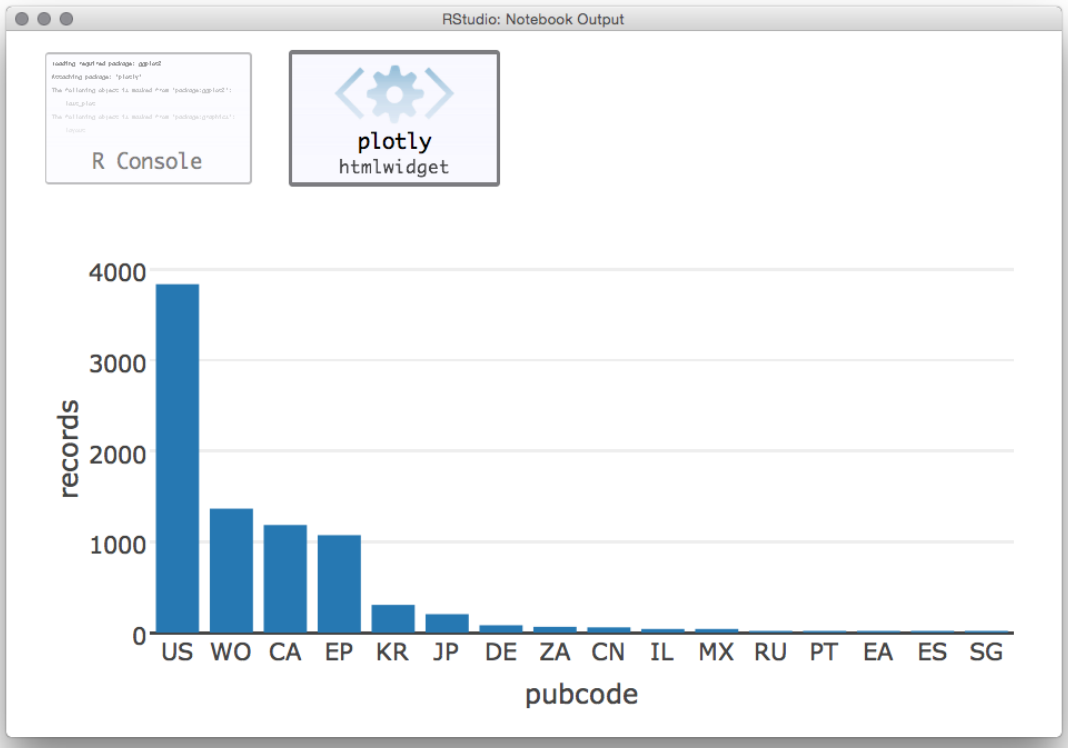
Para hacer esto necesitamos ejecutar algunos cálculos en los datos. Usaremos funciones de dplyry tidyrpara contar rápidamente la agrupación de datos por el código de publicación. Luego, agregaremos los datos a grupos discretos basados en las puntuaciones usando mutate()(para agregar una variable) y ntile()dividiremos los países en grupos según el número de registros (n) y agregaremos esto a la nueva variable llamada grupo. Finalmente, organizamos los datos en orden descendente según el número de registros.

```
library(dplyr)
```

```
  library(tidyr)
  total <- tally(group_by(pcy, pubcode)) %>%
    mutate(group = ntile(nn, 3)) %>%
    rename(records = nn) %>%
     arrange(desc(records))
   total
## # A tibble: 16 x 3
  ##    pubcode records group
  ##    <chr>     <int> <int>
  ##  1 US         3835     3
  ##  2 WO         1366     3
  ##  3 CA         1186     3
  ##  4 EP         1074     3
  ##  5 KR          307     3
  ##  6 JP          205     2
  ##  7 DE           83     2
  ##  8 ZA           66     2
  ##  9 CN           60     2
  ## 10 IL           29     2
  ## 11 MX           23     1
  ## 12 RU           10     1
  ## 13 PT            9     1
  ## 14 EA            4     1
  ## 15 ES            3     1
  ## 16 SG            2     1
```

Cuando vemos el total, ahora vemos que los países se han dividido en 3 grupos según el número de registros. Es poco probable que los grupos 1 y 2 proporcionen un gráfico significativo y, en particular, el grupo 1 podría eliminarse. Sin embargo, podríamos mostrar útil esta información como un gráfico de barras usando plot_lyy seleccionando type = "bar".

```
library(plotly)
  library(dplyr)
  total_bar <- plot_ly(total, x = pubcode , y = records, type = "bar")
total_bar
```



Habiendo dividido nuestros datos en tres grupos, ahora sería útil trazarlos por separado. Aquí enfrentamos el problema de que nuestros datos originales en pcy muestran valores por año, mientras que el total muestra el número total de registros y grupos. Primero necesitamos agregar los identificadores de grupo a la tabla pcy. Para hacer eso, modificaremos el total para eliminar el conteo de registros al recordsusar la dplyr select()función. Luego usaremos left_join()para unir las tablas totaly total_group. Tenga en cuenta que la función utilizará el campo compartido "código de publicación" para unirse.

```
library(dplyr)
  total_group <- select(total, pubcode, group)
   total_group
## # A tibble: 16 x 2
  ##    pubcode group
  ##    <chr>   <int>
  ##  1 US          3
  ##  2 WO          3
  ##  3 CA          3
  ##  4 EP          3
```

```
##  5 KR          3
##  6 JP          2
##  7 DE          2
##  8 ZA          2
##  9 CN          2
## 10 IL          2
## 11 MX          1
## 12 RU          1
## 13 PT          1
## 14 EA          1
## 15 ES          1
## 16 SG          1
```

Luego unimos las dos tablas y cambiamos el nombre na recordspara graficar.

```
library(dplyr)
  total_grouped <- left_join(pcy, total_group) %>%
     rename(records = n)
## Joining, by = "pubcode"
total_grouped
## # A tibble: 223 x 5
  ##    pubcountry pubcode pubyear records group
  ##    <chr>      <chr>     <int>   <int> <int>
  ##  1 Canada     CA         1990      19     3
  ##  2 Canada     CA         1991      49     3
  ##  3 Canada     CA         1992      66     3
  ##  4 Canada     CA         1993      59     3
  ##  5 Canada     CA         1994      50     3
  ##  6 Canada     CA         1995      39     3
  ##  7 Canada     CA         1996      36     3
  ##  8 Canada     CA         1997      45     3
  ##  9 Canada     CA         1998      46     3
  ## 10 Canada     CA         1999      47     3
  ## # ... with 213 more rows
```

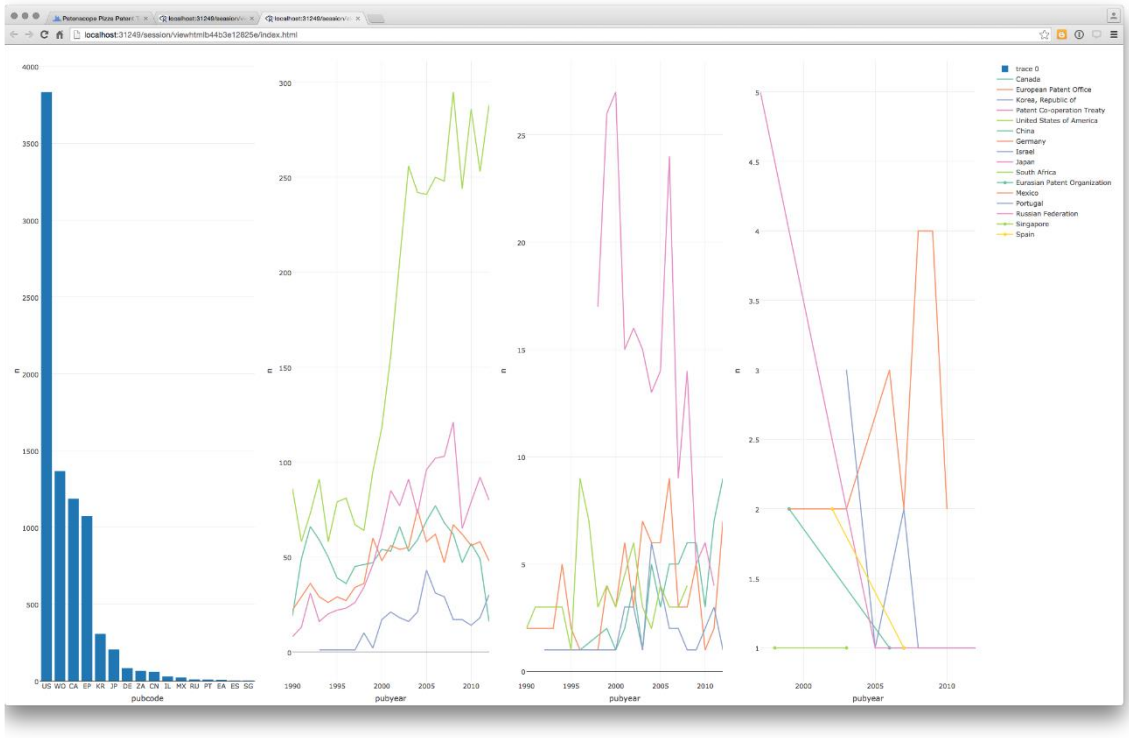
El siguiente paso es generar un conjunto de tres gráficos correspondientes a nuestros tres grupos. Los llamaremos pizza3, pizza2 y pizza1 y usaremos el nombre completo del país de publicación pubcountrycomo el color de las líneas.

```
library(plotly)


library(dplyr)


  pizza3 <- filter(total_grouped, group == 3) %>%
   plot_ly(x = pubyear, y = records, color = pubcountry,
  type = "lines", mode = "lines")


pizza2 <- filter(total_grouped, group == 2) %>%
  plot_ly(x = pubyear, y = records, color = pubcountry,
 type = "lines", mode = "lines")
pizza1 <- filter(total_grouped, group == 1) %>%
    plot_ly(x = pubyear, y = records, color = pubcountry,
   type = "lines", markers = "lines")
```

Ahora tenemos un total de cuatro borradores de parcelas, barra total y pizza 3 a 1 para nuestros grupos. Plotly nos permitirá mostrar las parcelas lado a lado. Tenga en cuenta que esto puede crear una visualización bastante crujiente en RStudio y se ve mejor seleccionando el show in new windowbotón pequeño en el RStudio Viewer.

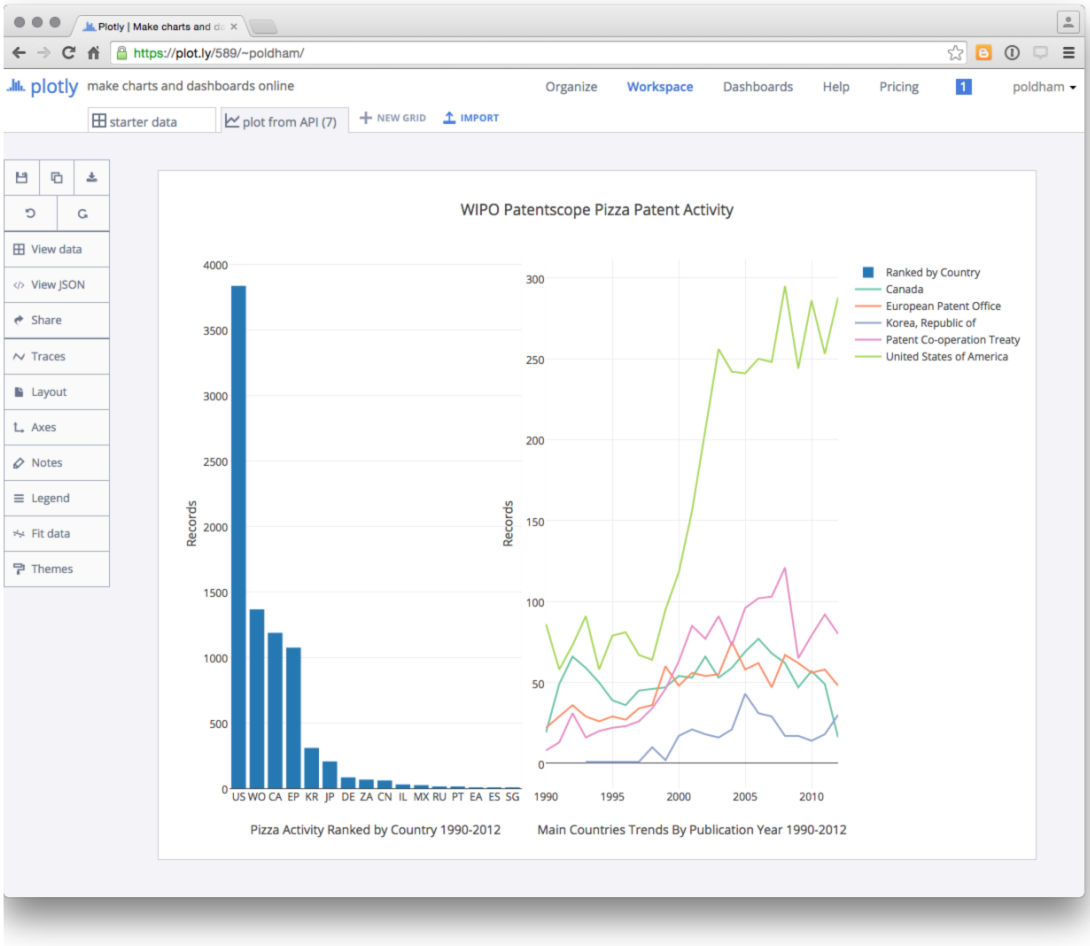
```
library(plotly)
  sub <- subplot(total_bar, pizza3, pizza2, pizza1)
  sub
  plotly_POST(sub)
```

Ahora verá una imagen que se parece mucho a esto.



La figura no revela una tendencia coherente para los países en el Grupo 1 a la derecha y tiene sentido eliminar estos datos. El grupo 2 es potencialmente más interesante, pero las cifras generales bajas y los picos de datos para Japón sugieren una actividad muy baja y una falta de datos completos. Además, lo ideal sería que quisiéramos asignar diferentes colores a los diferentes nombres en nuestros paneles de tendencias (probablemente mediante la asignación de diferentes paletas), lo que podría llevar un tiempo considerable en relación con las ganancias en términos de mostrar datos de baja frecuencia. Dejaremos que el gráfico de barras haga ese trabajo y terminaremos con un simple gráfico de dos parcelas para enviar en plotlylínea.

```
library(plotly)
  sub1 <- subplot(total_bar, pizza3)
  plotly_POST(sub1)
```



Entonces es fácil editar las etiquetas y hacer los ajustes finales en línea.

También podemos compartir el gráfico a través de las redes sociales, descargar los datos o editar el gráfico. Tenga en cuenta que la configuración predeterminada para un gráfico enviado a través de la API parece ser pública (sin una forma obvia de cambiar eso).

Aquí es donde la importancia potencial de Plotly como herramienta para compartir datos y gráficos se hace evidente. Es una herramienta poderosa. Las actualizaciones recientes del paquete R y la introducción de paneles de control demuestran las mejoras continuas de este nuevo servicio.

11.7 Round Up

En este capítulo, proporcionamos una breve introducción a Plotly para ayudarlo a comenzar a utilizar esta herramienta para el análisis de patentes. Plotly proporciona gráficos visualmente atractivos e interactivos que se pueden compartir fácilmente

con colegas, pegarlos en sitios web y compartirlos públicamente. La disponibilidad de API también es una característica clave de Plotly para aquellos que trabajan en Python, R u otros entornos programáticos.

Sin embargo, Plotly también puede ser confuso. Por ejemplo, nos resultó difícil entender por qué determinados conjuntos de datos no se cargarían correctamente (cuando se pueden leer fácilmente en Tableau). También nos resultó difícil entender el formato en el que los datos debían estar para realizar un trazado correcto. Por lo tanto, Plotly puede ser algo frustrante aunque tiene un potencial muy considerable para compartir gráficos atractivos. La reciente incorporación de paneles es también un desarrollo prometedor. Finalmente, para los usuarios de R, el plotlypaquete ahora se integra estrechamente con el ggplot2paquete muy popular a través de la función ggplotly () que permite la creación de ggplot2gráficos interactivos .

En este capítulo, solo hemos tratado el potencial de Plotly como una poderosa herramienta gratuita para crear gráficos interactivos. Otros tipos de parcelas que vale la pena explorar incluyen mapas de burbujas, mapas de contorno y mapas de calor. Para experimentar por ti mismo prueba los tutoriales de Plotly .

Capítulo 12 Infografía de patentes con R

En este capítulo usaremos RStudio para preparar datos de patentes para visualización en una infografía utilizando herramientas de software en línea.

Las infografías son una forma popular de presentar datos de una manera que es fácil de entender para un lector sin leer un informe largo. Las infografías son adecuadas para presentar resúmenes de datos con mensajes simples sobre hallazgos clave. Una buena infografía puede animar a la audiencia a leer un informe detallado y es una herramienta para relacionarse con las audiencias durante las presentaciones de los resultados de la investigación de patentes.

Algunas oficinas de patentes ya han estado creando infografías como parte de sus informes a los encargados de formular políticas y otros clientes. El Instituto Nacional de Propiedad Industrial (INPI) en Brasil produce un radar de tecnología de dos páginas (Radar Tecnológico) que consiste en cuadros y mapas que resumen brevemente investigaciones más detalladas sobre temas como la nanotecnología en la gestión de residuos . Los Informes de Patentes de la OMPI sobre el paisaje , que profundizan en la actividad de patentes para un área en particular, están acompañados por infografías de una página que han demostrado ser muy populares, como la infografía que acompaña a un informe reciente sobre dispositivos de asistencia .

Un número creciente de compañías están ofreciendo servicios de software de infografía en línea como infogr.am , easel.ly piktochart.com , canva.com o venngage.com para mencionar solo una selección de las ofertas disponibles. El sitio web de Cool Infographicsproporciona una visión general útil de las herramientas disponibles.

Una característica de muchos de estos servicios es que se basan en un modelo freemium. La creación de gráficos es gratuita, pero la capacidad de exportar archivos y los formatos disponibles para exportar su obra maestra (por ejemplo, alta resolución o .pdf) a menudo dependen de la actualización a una cuenta mensual a precios variables. En este capítulo, probamos drive infogr.am como un servicio amigable con los gráficos, aunque con opciones de exportación que dependen de una cuenta pagada.

Este capítulo está dividido en dos secciones.

1. En la parte 1 nos centramos en el uso de rstudio para preparar los datos de patentes para la visualización en el software de infografía usando los dplyr,

tidyry stringrpaquetes. Esto implica tratar problemas comunes con datos de patentes, como campos concatenados, espacios en blanco y la creación de conteos de campos de datos.

2. En la parte 2 producimos una infografía a partir de los datos utilizando infogr.am .

12.1 Primeros pasos

Para empezar, necesitamos asegurarnos de que RStudio y R para su sistema operativo estén instalados siguiendo las instrucciones en el sitio web de RStudio aquí . No olvide seguir el enlace para instalartambién R para su sistema operativo.

Cuando se trabaja en RStudio, es una buena práctica trabajar con proyectos. Esto mantendrá todos los archivos para un proyecto en la misma carpeta. Para crear un proyecto, vaya a Archivo, Nuevo proyecto y cree un proyecto. Llama al proyecto algo así como infografía. Cualquier archivo que cree y guarde para el proyecto ahora aparecerá en la pestaña Archivos en RStudio.

R trabaja con paquetes (bibliotecas) y hay alrededor de 7,490 de ellos para una amplia gama de propósitos. Vamos a utilizar sólo algunos de ellos. Para instalar un paquete utilizamos lo siguiente. Copie y pegue el código en la Consola y presione enter.

```
install.packages("tidyverse")  # the group of packages you will need
```

Los paquetes también se pueden instalar seleccionando la pestaña Paquetes y escribiendo el nombre del paquete.

Para cargar los paquetes (bibliotecas) use lo siguiente o marque la casilla de verificación en el panel Paquetes.

```
library(readr)
  library(dplyr)
  library(tidyr)
  library(stringr)
  library(ggplot2)
```

Ahora estamos listos para irnos.

12.2 Cargar un archivo .csv usandoreadr

Trabajaremos con el pizza_medium_cleanconjunto de datos en el repositorio de Github Manual en línea . Si descarga un archivo manualmente, recuerde hacer clic en el nombre del archivo y seleccione Rawdescargar el archivo real.

Podemos usar la función fácil de usar read_csv()del readrpaquete para leer rápidamente nuestros datos de pizza directamente desde el repositorio de Github. Tenga rawen cuenta al principio del nombre de archivo.

```
library(readr)
  pizza <- read_csv
("https://github.com/wipo-analytics/opensource-patent-analytics/
blob/master/2_datasets/pizza_medium_clean/pizza.csv?raw=true")
```

readrmostrará una advertencia para el archivo que surge de sus esfuerzos para analizar las fechas de publicación en la importación. Ignoraremos esto ya que no usaremos este campo.

Como alternativa a la importación directamente desde Github, descargue el archivo y utilice RStudio File > Import Dataset > From .csv. Si experimenta problemas con la importación directa de un archivo, el enfoque Archivo> Importar conjunto de datos le brindará un rango de controles fáciles de usar para resolver esto (por ejemplo, donde .csv es en realidad un archivo separado por tabulaciones).

12.3 Visualización de datos

Podemos ver los datos de varias maneras.

1. En la consola:

```
pizza
## # A tibble: 9,996 × 31
  ## applicants_cleaned
  ## <chr>
  ## 1 <NA>
 ## 2 Ventimeglia Jamie Joseph; Ventimeglia Joel Michael; Ventimeglia
Thomas Jose
  ## 3 Cordova Robert; Martinez Eduardo
  ## 4 Lazarillo De Tormes S L
 ## 5 <NA>
 ## 6 Depoortere, Thomas
 ## 7 Frisco Findus Ag
```

```
## 8 Bicycle Tools Incorporated
## 9 Castiglioni, Carlo
## 10 <NA>
## # ... with 9,986 more rows, and 30 more variables:
## #   applicants_cleaned_type <chr>, applicants_organisations <chr>,
## #   applicants_original <chr>,
     inventors_cleaned <chr>,
## #   inventors_original <chr>,
 ipc_class <chr>, ipc_codes <chr>,
## #   ipc_names <chr>,
      ipc_original <chr>, ipc_subclass_codes <chr>,
## #   ipc_subclass_detail <chr>,
ipc_subclass_names <chr>,
## #   priority_country_code <chr>, priority_country_code_names <chr>,
 ## #   priority_data_original <chr>,
  priority_date <chr>,
 ## #   publication_country_code <chr>,
  publication_country_name <chr>,
## # publication_date <chr>,
 publication_date_original <chr>,
## #   publication_day <int>, publication_month <int>,
## # publication_number <chr>, publication_number_espacenet_links <chr>,
## #   publication_year <int>,
title_cleaned <chr>, title_nlp_cleaned <chr>,
## #   title_nlp_multiword_phrases <chr>, title_nlp_raw <chr>,
## #   title_original <chr>
```

2. En Entorno haz clic en la flecha azul para ver en el entorno. Sigue haciendo clic para abrir una nueva ventana con los datos.
3. Utilice el View()comando (para data.frames y tablas)

```
View(pizza)
```

Si es posible, use el comando o entorno View (). La dificultad con la consola es que grandes cantidades de datos simplemente se transmitirán al pasado.

12.4 Identificación de tipos de objetos

A menudo queremos saber con qué tipo de objeto estamos trabajando y más detalles sobre el objeto para saber qué hacer más adelante. Estos son algunos de los comandos más comunes para obtener información sobre objetos.

```
class(pizza)
## type of object
names(pizza)
## names of variables
str(pizza)
## structure of object
dim(pizza)
## dimensions of the object
```

El comando más útil en esta lista es str()porque nos permite acceder a la estructura del objeto y ver su tipo.

```
str(pizza, max.level = 1)
```

str()Es particularmente útil porque podemos ver los nombres de los campos (vectores) y su tipo. La mayoría de los datos de patentes es un vector de caracteres con fechas que forman números enteros.

12.5 Trabajando con datos

A menudo querremos seleccionar aspectos de nuestros datos para enfocarnos en un conjunto específico de columnas o para crear un gráfico. Es posible que también queramos agregar información, especialmente conteos numéricos.

El dplyrpaquete proporciona un conjunto de funciones muy útiles para seleccionar, agregar y contar datos. Los paquetes tidyry stringrson paquetes hermanos que contienen una gama de otras funciones útiles para trabajar con nuestros datos. Hemos cubierto algunos de estos en otros capítulos sobre gráficas con R, pero los analizaremos rápidamente y luego los agruparemos en una función que podemos usar en nuestro conjunto de datos.

12.5.1 Seleccionar

En este caso, comenzaremos utilizando la select()función para limitar los datos a columnas específicas. Podemos hacer esto usando sus nombres o su posición numérica (mejor para un gran número de columnas, por ejemplo, 1:31). En dplyr, a diferencia de la mayoría de los paquetes R, las columnas de caracteres existentes no requieren "".

```
library(dplyr)
pizza_number <- select(pizza, publication_number, publication_year)
```

Ahora tenemos un nuevo data.frame que contiene dos columnas. Uno con el año y otro con el número de publicación. Tenga en cuenta que hemos creado un nuevo objeto llamado pizza_number usando <-y que, después de select()que hayamos nombrado nuestros datos originales y las columnas que deseamos. Una característica fundamental de la selección es que eliminará las columnas que no nombramos. Por lo tanto, es mejor crear un nuevo objeto utilizando <-si desea conservar sus datos originales para su trabajo posterior.

12.5.2 Agregando datos conmutate()

mutate()es una dplyrfunción que nos permite agregar datos basados en datos existentes en nuestro marco de datos, por ejemplo, para realizar un cálculo. En el caso de datos de patentes, normalmente carecemos de un campo numérico para usar en los recuentos. Sin embargo, podemos asignar un valor a nuestro campo de publicación utilizando sum () y el número 1 de la siguiente manera.

```
library(dplyr)
pizza_number <- mutate(pizza_number, n = sum(publication_number = 1))
```

Cuando vemos pizza_number, ahora tenemos un valor de 1 en la columna npara cada número de publicación. Tenga en cuenta que en los datos de patentes puede aparecer un número de prioridad, solicitud, publicación o familia varias veces y desearíamos reducir el conjunto de datos a registros distintos. Para eso usaríamos n_distinct (pizza_number$publication_number)

desde dplyro unique (pizza_number$publication_number) desde la base R. Ya que los números de publicación son únicos, podemos proceder.

12.5.3 Contando datos utilizandobcount()

En este momento, tenemos varias instancias del mismo año (donde se produce una publicación de patente en ese año). Ahora queremos calcular cuántos de nuestros documentos se publicaron en cada año. Para ello utilizaremos la dplyrfunción count(). Usaremos la publicación año y agregaremos wt =(para ponderar) ncomo el valor a contar.

```
library(dplyr)
pizza_total <- count(pizza_number, publication_year, wt = n)
pizza_total
```

```
## # A tibble: 58 × 2
    ##    publication_year    nn
    ##               <int> <dbl>
    ## 1              1940     1
    ## 2              1954     1
    ## 3              1956     1
    ## 4              1957     1
    ## 5              1959     1
    ## 6              1962     1
    ## 7              1964     2
    ## 8              1966     1
    ## 9              1967     1
    ## 10             1968     8
    ## # ... with 48 more rows
```

Cuando ahora examinemos pizza_total, veremos el año de publicación y un valor sumado para los registros de ese año.

Esto plantea la cuestión de cómo sabemos que R ha calculado el recuento correctamente. Ya sabemos que hay 9996 registros en el conjunto de datos de pizza. Para verificar que nuestro conteo sea correcto, simplemente podemos usar la suma y seleccionar la columna que queremos sumar usando $.

```
library(dplyr)
        sum(pizza_total$nn)
```

```
## [1] 9996
```

Entonces, todo está bien y podemos seguir adelante. El $signo es una de las principales formas de subconjunto para indicar a R que queremos trabajar con una columna específica (las otras son "[" y "[[").

12.5.4 Renombrar un campo conrename()

A continuación usaremos rename()desde dplyrpara renombrar los campos. Tenga en cuenta que comprender qué campo requiere comillas puede requerir cierto esfuerzo. En este caso, cambiar el nombre del vector de caracteres publicación_year como "pubyear" requiere comillas, mientras que cambiar el nombre del vector numérico "n" no.

```
library(dplyr)
```

```
pizza_total <- rename(pizza_total, pubyear = publication_year,
 publications = nn) %>%
    print()
```

```
## # A tibble: 58 × 2
    ##    pubyear publications
    ##      <int>        <dbl>
    ## 1     1940            1
    ## 2     1954            1
    ## 3     1956            1
     ## 4     1957            1
     ## 5     1959            1
     ## 6     1962            1
     ## 7     1964            2
     ## 8     1966            1
     ## 9     1967            1
     ## 10    1968            8
     ## # ... with 48 more rows
```

12.5.5 Hacer una gráfica rápida conqplot()

Usando la qplot()función en ggplot2ahora podemos dibujar un gráfico de líneas rápido. Tenga en cuenta que qplot () es inusual en R porque los datos (pizza_total) aparecen después de las coordenadas. Especificaremos que queremos usar una línea geom =(si geom se deja fuera, será un diagrama de dispersión). Esto nos dará una idea de cómo se vería nuestra trama en nuestra infografía y las acciones que podríamos querer tomar en los datos.

```
library(ggplot2)
 qplot(x = pubyear,
  y = publications, data = pizza_total,
   geom = "line")
```

La trama revela un acantilado de datos en los últimos años. Esto normalmente refleja una falta de datos durante los últimos 2-3 años a medida que los documentos recientes se alimentan a través del sistema en el camino hacia la publicación.

Es una buena idea eliminar el acantilado de datos cortando los datos dos o tres años antes del presente. En algunos casos, dos años es suficiente, pero 3 años es una buena regla general.

También tenemos una larga cola de datos con datos limitados desde 1940 hasta finales de los años setenta. Dependiendo de nuestros propósitos con el análisis, podríamos querer mantener estos datos (para el análisis histórico) o enfocarnos en un período más reciente.

Limitaremos nuestros datos a valores específicos utilizando la dplyrfunción filter().

12.5.6 Filtrar datos utilizando filter()

A diferencia de lo select()que funciona con columnas, filter()en dplyrtrabajos con filas. En este caso, necesitamos filtrar los valores en la columna de pubyear. Para eliminar los datos anteriores a 1990, utilizaremos el operador mayor o igual que >=el de la columna Pubyear y utilizaremos el <=operador menor o igual que en los valores posteriores a 2012.

Una fuerza de filter()en dplyres que es fácil de filtrar en múltiples valores en la misma expresión (a diferencia de la función de filtro muy similar en la base de R). El uso de filter()también eliminará los 30 registros en los que el año se registra como NA (No disponible). Escribiremos este archivo en el disco usando el simple write_csv()de readr. Para usarlo write_csv(), primero asignamos un nombre a nuestros datos (pizza_total) y luego proporcionamos un nombre de archivo con la extensión .csv. En este caso y en otros ejemplos a continuación, hemos utilizado un nombre de archivo descriptivo teniendo en cuenta que los sistemas de Windows tienen limitaciones en la longitud y el tipo de caracteres que se pueden usar en los nombres de archivo.

```
library(dplyr)
library(readr)
pizza_total <- filter(pizza_total, pubyear >= 1990, pubyear <= 2012)
write_csv(pizza_total, "pizza_total_1990_2012.csv")
pizza_total
## # A tibble: 23 × 2
    ##    pubyear publications
    ##      <int>        <dbl>
    ## 1     1990          139
    ## 2     1991          154
```

```
## 3     1992          212
## 4     1993          201
## 5     1994          162
## 6     1995          173
## 7     1996          180
## 8     1997          186
## 9     1998          212
## 10    1999          290
## # ... with 13 more rows
```

Cuando imprimamos pizza_total en la consola, veremos que los datos ahora cubren el período 1990-2012. Cuando se usa filter()en valores de esta manera, es importante recordar aplicar este filtro a cualquier operación posterior en los datos (como los solicitantes) para que coincida con el mismo período de datos.

Para ver nuestro archivo .csv podemos dirigirnos a la pestaña Archivos y, asumiendo que hemos creado un proyecto, el archivo ahora aparecerá en la lista de archivos del proyecto. Al hacer clic en el nombre del archivo se mostrarán los datos sin formato sin formato en RStudio.

12.6 Simplificar el código con tuberías.%>%

Hasta ahora hemos manejado el código una línea a la vez. Pero, una de las grandes fortalezas de usar un lenguaje de programación es que podemos ejecutar varias líneas de código juntas. Hay dos formas básicas en que podemos hacer esto.

Crearemos un nuevo objeto temporal dfpara demostrar esto.

1. La forma estandar

```
library(dplyr)
library(ggplot2)
df <- select(pizza, publication_number, publication_year)
df <- mutate(df, n = sum(publication_number = 1))
df <- count(df, publication_year, wt = n)
df <- rename(df, pubyear = publication_year, publications = nn)
df <- filter(df, pubyear >= 1990, pubyear <= 2012)
qplot(x = pubyear, y = publications, data = df, geom = "line")
```

El código que acabamos de crear es de seis líneas. Si seleccionamos todo este código y lo ejecutamos de una sola vez, producirá nuestro gráfico.

Una característica de este código es que cada vez que ejecutamos una función en el total del objeto, la nombramos al inicio de cada función (por ejemplo, mutate (df ...)) y luego sobrescribimos el objeto.

Podemos ahorrar bastante escritura y reducir la complejidad del código utilizando el operador de tubería introducido por el magrittrpaquete y luego adoptado en los paquetes de ordenación y ordenación de datos de Hadley Wickham.

2. Utilizando tuberías %>%

Las tuberías son ahora una forma muy popular de escribir código R porque simplifican la escritura del código R y lo aceleran. La tubería más popular es lo %>%que significa "esto" y luego "eso". En este caso, vamos a crear un nuevo objeto temporal df1aplicando primero a pizza, luego mutamos, contamos, renombramos y filtramos. Tenga en cuenta que solo asignamos un nombre a nuestro conjunto de datos una vez (in select()) y no tenemos que seguir sobrescribiendo el objeto.

```
library(dplyr)
library(ggplot2)
df1 <- select(pizza, publication_number, publication_year)
 %>% mutate(n = sum(publication_number = 1)) %>%

count(publication_year, wt = n)
 %>% rename(pubyear = publication_year, publications = nn) %>%

filter(pubyear >= 1990, pubyear <= 2012)
 %>% qplot(x = pubyear, y = publications,
data = ., geom = "line") %>% print()
```

En el código estándar escribimos df nueve veces para llegar al mismo resultado. Usando tuberías escribimos df1 una vez. De mayor importancia es que el uso de tuberías simplifica la estructura del código R mediante la introducción de una lógica básica de "esto" y luego de "eso" que facilita su comprensión.

Un punto a tener en cuenta sobre este código es que qplot()nos obliga a asignar un nombre a nuestros datos (en este caso df1). Sin embargo, en df1realidad es la salida final del código y no existe como un objeto de entrada antes de que se ejecute la línea final. Por lo tanto, si tratamos de utilizar data = df1en qplot()recibiremos un mensaje de error. La forma de evitar esto es utilizarlo .en lugar de nuestro objeto de datos. De esa manera qplot()sabremos que queremos graficar las salidas del código anterior. Finalmente, necesitamos agregar una llamada explícita print()para mostrar el gráfico (sin esto, el código funcionará pero no veremos el gráfico).

Si ahora inspeccionamos la estructura del objeto df1 (usando str(df1)) en la consola, será una lista. La razón de esto es que es un objeto con componentes mixtos, que incluye un data.frame con nuestros datos más datos adicionales que establecen el contenido de la gráfica. Como no hay un enlace directo entre R y nuestro software de infografía, esto nos creará problemas más adelante porque el software de infografía no sabrá cómo interpretar el objeto de la lista. Por lo tanto, generalmente es una buena idea usar un data.frame directo al excluir la llamada qploty agregarla más tarde cuando sea necesario de la siguiente manera.

```
library(dplyr)
library(ggplot2)
df2 <- select(pizza, publication_number, publication_year)
%>% mutate(n = sum(publication_number = 1)) %>%
      count(publication_year, wt = n) %>% rename
  (pubyear = publication_year, publications = nn) %>%
filter(pubyear >= 1990, pubyear <= 2012) %>% print()
## # A tibble: 23 × 2
  ##    pubyear publications
  ##      <int>        <dbl>
  ## 1     1990          139
  ## 2     1991          154
  ## 3     1992          212
  ## 4     1993          201
  ## 5     1994          162
  ## 6     1995          173
  ## 7     1996          180
  ## 8     1997          186
  ## 9     1998          212
  ## 10    1999          290
  ## # ... with 13 more rows
```

Tenga en cuenta que, en este caso, el único cambio es que debemos incluir explícitamente la referencia al marco de datos df2 como el argumento de datos en la llamada a qplot().

```
library(ggplot2)
 qplot(x = pubyear, y = publications, data = df2, geom = "line")
```

12.7 Armonización de datos.

Un desafío al crear varias tablas a partir de un conjunto de datos de línea de base es hacer un seguimiento de los conjuntos de datos. En este momento tenemos dos objetos básicos con los que trabajaremos:

1. pizza - nuestro conjunto de datos en bruto
2. pizza_total- Creado vía pizza_number limitada a 1990_2012.

En el resto del capítulo, desearemos crear algunos conjuntos de datos adicionales a partir de nuestro conjunto de datos de pizza. Estos son:

1. Tendencias del país
2. Solicitantes
3. Clase de Clasificación Internacional de Patentes (IPC)
4. Frases
5. Google
6. Google IPC
7. Frases de google

Debemos asegurarnos de que todos los datos que generemos a partir de nuestro conjunto de datos sin procesar coincidan con el período para el pizza_totalconjunto de datos. Si no lo hacemos, existe el riesgo de que generemos subdatasets con recuentos para el conjunto de datos de pizza sin procesar.

Para manejar esto usaremos filter()para crear un nuevo conjunto de datos de línea de base con un nombre no ambiguo.

```
library(dplyr)
pizza_1990_2012
 <- rename(pizza, pubyear = publication_year) %>% filter(pubyear >=
1990, pubyear <= 2012)


pizza_1990_2012
## # A tibble: 8,262 × 31
```

```
## applicants_cleaned applicants_cleaned_type
## <chr> <chr>
## 1 <NA> People
## 2 Lazarillo De Tormes S L Corporate
## 3 <NA> People
## 4 Depoortere, Thomas People
## 5 Frisco Findus Ag Corporate
## 6 Bicycle Tools Incorporated Corporate
## 7 Castiglioni, Carlo People
## 8 <NA> People
## 9 Bujalski, Wlodzimierz People
## 10 Ehrno Flexible A/S; Stergaard, Ole Corporate; People
## # ... with 8,252 more rows, and 29 more variables:
## # applicants_organisations <chr>, applicants_original <chr>,
## # inventors_cleaned <chr>, inventors_original <chr>,
ipc_class <chr>,
## # ipc_codes <chr>, ipc_names <chr>, ipc_original <chr>,
## # ipc_subclass_codes <chr>, ipc_subclass_detail <chr>,
## # ipc_subclass_names <chr>, priority_country_code <chr>,
## # priority_country_code_names <chr>, priority_data_original <chr>,
## # priority_date <chr>, publication_country_code <chr>,
## # publication_country_name <chr>, publication_date <chr>,
## # publication_date_original <chr>, publication_day <int>,
## # publication_month <int>, publication_number <chr>,
## # publication_number_espacenet_links <chr>, pubyear <int>,
## # title_cleaned <chr>, title_nlp_cleaned <chr>,
## # title_nlp_multiword_phrases <chr>, title_nlp_raw <chr>,
## # title_original <chr>
```

En este caso, comenzamos con una llamada para rename()hacer que esto sea coherente con nuestra tabla pizza_total y luego usamos una tubería para filtrar los datos del año. Tenga en cuenta que al filtrar datos sin procesar en un conjunto de valores, es importante inspeccionarlos primero para verificar que el campo esté limpio (por ejemplo, no concatenado). Si, por alguna razón, sus datos están concatenados (lo que ocurre bastante con los datos de patentes), realice una búsqueda ?tidyr::separate_rows.

Ahora estamos en condiciones de crear nuestra tabla de tendencias de país.

12.8 Tendencias de país utilizandospread()

Hay dos formatos de datos básicos: largo y ancho. Nuestro conjunto de datos de pizza está en formato largo porque cada columna es una variable (por ejemplo publication_country) y cada fila publication_countrycontiene un nombre de país. Este es el formato de datos más común y útil.

Sin embargo, en algunos casos, como infogr.amnuestro software de visualización, esperará que los datos estén en formato ancho. En este caso, cada nombre de país se convertiría en una variable (nombre de columna) con los años que forman las filas y el número de registros por año de las observaciones. La clave para esto es la tidyr()función spread().

Como antes, comenzaremos utilizando select()para crear una tabla con los campos que deseamos. Luego lo usaremos mutate()para agregar un campo numérico y luego contar esos datos. Para ilustrar el proceso ejecute este código (no crearemos un objeto).

```
library(dplyr)
select(pizza_1990_2012, publication_country_name, publication_number,
 pubyear) %>%

    mutate(n = sum(publication_number = 1))
 %>% count(publication_country_name,

    pubyear, wt = n) %>% print()

## Source: local data frame [223 x 3]
  ## Groups: publication_country_name [?]
  ##   ##    publication_country_name pubyear    nn
  ## <chr>   <int> <dbl>
  ## 1 Canada    1990    19
  ## 2 Canada    1991    49
  ## 3 Canada    1992    66
  ## 4 Canada    1993    59
  ## 5 Canada    1994    50
  ## 6 Canada    1995    39
  ## 7 Canada    1996    36
  ## 8 Canada    1997    45
```

```
  ## 9 Canada    1998    46
  ## 10 Canada    1999    47
  ## # ... with 213 more rows
```

Cuando ejecutemos este código veremos los resultados en formato largo. Ahora queremos tomar nuestra publication_country_namecolumna y extenderla para formar columnas con nnlos valores.

En el uso de la propagación, tenga en cuenta que toma un argumento de datos (pizza_1990_2012), una clave (publication_country_name) y una columna de valor (nn) (creada desde count()). Estamos utilizando tuberías, por lo que los datos solo deben mencionarse en la primera línea. Para argumentos adicionales ver ?spread().

```
library(dplyr)
  library(tidyr)
  country_totals
 <- select(pizza_1990_2012, publication_country_name, publication_number,
  pubyear) %>%
   mutate(n = sum(publication_number = 1)) %>%
   count(publication_country_name, pubyear, wt = n)
      %>% # note n
   spread(publication_country_name, nn) # note double nn
 country_totals
## # A tibble: 23 × 17
  ## pubyear Canada China `Eurasian Patent Organization`
## *    <int>  <dbl> <dbl> <dbl>
    ## 1     1990     19    NA NA
       ## 2     1991     49    NA NA
       ## 3     1992     66    NA NA
       ## 4     1993     59    NA NA
       ## 5     1994     50    NA NA
       ## 6     1995     39    NA NA
       ## 7     1996     36     1 NA
       ## 8     1997     45    NA NA
       ## 9     1998     46    NA NA
       ## 10    1999     47     2 2
       ## # ... with 13 more rows, and 13
        more variables: `European Patent
        ## #   Office` <dbl>, Germany <dbl>
```

```
        , Israel <dbl>, Japan <dbl>
        , `Korea,
        ## #   Republic of` <dbl>
        , Mexico <dbl>
        , `Patent Co-operation Treaty` <dbl>,
        ## #   Portugal <dbl>
        , `Russian Federation` <dbl>
        , Singapore <dbl>, `South
        ## #   Africa` <dbl>
        , Spain <dbl>
        , `United States of America` <dbl>
```

Ahora tenemos datos en formato ancho.

En algunos casos, como infogr.am, el software de visualización puede esperar que los nombres de los países sean el nombre de las filas y que los nombres de las columnas sean años. Podemos modificar nuestra llamada para spread()reemplazar publication_country_namecon pubyear. Luego, escribiremos los datos en el disco para utilizarlos en nuestra infografía.

```
library(dplyr)
 library(readr)
  country_totals <- select(pizza_1990_2012,
  publication_country_name, publication_number, pubyear) %>%
   mutate(n = sum(publication_number = 1)) %>%
   count(publication_country_name, pubyear, wt = n) %>%
  # note n
   spread(pubyear, nn) # note nn country_totals
## Source: local data frame [16 x 24]
  ## Groups: publication_country_name [16]
  ##
  ##publication_country_name
`1990` `1991` `1992` `1993`    `1994` `1995`
   ## * <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
  ## 1 Canada 19 49 66 59 50 39
  ## 2 China NA NA NA NA NA NA
  ## 3  Eurasian Patent Organization NA NA NA NA NA NA
  ## 4 European Patent Office 22 29 36 29 26 29
  ## 5 Germany 2 2 2 2 5 2
```

```
  ## 6 Israel NA NA 1 NA NA 1
  ## 7 Japan     NA     NA     NA     NA     NA     NA
  ## 8 Korea, Republic of NA     NA     NA      1     NA     NA
  ## 9 Mexico     NA     NA     NA     NA     NA     NA
  ## 10 Patent Co-operation Treaty 8 13 31 16 20 22
  ## 11 Portugal NA     NA     NA     NA     NA     NA
  ## 12 Russian Federation NA NA NA     NA     NA     NA
  ## 13 Singapore     NA NA     NA     NA     NA     NA
  ## 14 South Africa      2      3      3      3      3      1
  ## 15 Spain     NA     NA     NA     NA     NA     NA
  ## 16 United States of America 86 58 73 91 58     79
  ## # ... with 17 more variables:
  `1996` <dbl>,
  `1997` <dbl>,
  `1998` <dbl>,
  ## #   `1999` <dbl>,
  `2000` <dbl>,
  `2001` <dbl>,
  `2002` <dbl>,
  `2003` <dbl>,
  ## #   `2004` <dbl>,
  `2005` <dbl>,
  `2006` <dbl>,
  `2007` <dbl>,
  `2008` <dbl>,
  ## #   `2009` <dbl>,
  `2010` <dbl>,
  `2011` <dbl>,
  `2012` <dbl>
write_csv(country_totals, "pizza_country_1990_2012.csv")
```

Para restaurar los datos a un formato largo, tendríamos que utilizarlos gather()como contrapartida spread(). gather()toma un conjunto de datos, una clave para el nombre de la columna en la que queremos reunir los países, un valor para el recuento numérico (en este caso n) y, finalmente, las posiciones de las columnas en las que se debe reunir. Tenga en cuenta que necesitamos busque las posiciones de columna en country_totals(por ejemplo, usando View()) o cuente las columnas usando ncol(country_totals).

```
library(dplyr)
gather(country_totals, year, n, 2:24) %>% print()

## Source: local data frame [368 x 3]
## Groups: publication_country_name [16]
##   ##        publication_country_name  year     n
## <chr> <chr> <dbl>
## 1                         Canada  1990    19
## 2                          China  1990    NA
## 3  Eurasian Patent Organization  1990    NA
## 4        European Patent Office  1990    22
## 5                        Germany  1990     2
## 6                         Israel  1990    NA
## 7                          Japan  1990    NA
## 8            Korea, Republic of  1990    NA
## 9                         Mexico  1990    NA
## 10   Patent Co-operation Treaty  1990     8
## # ... with 358 more rows
```

La combinación de difusión y recopilación funciona realmente bien para preparar los datos en formatos que otros programas esperan. Sin embargo, uno de los principales problemas que encontramos con los datos de patentes es que nuestros datos no están ordenados porque varios campos están concatenados.

12.9 Ordenando datos - Separando y recolectando

En los datos de patentes, a menudo vemos campos concatenados con un separador (normalmente a ;). Estos son, por lo general, nombres de solicitantes, nombres de inventores, códigos de Clasificación Internacional de Patentes (CIP) o números de documentos (números de prioridad, números de familia). Necesitamos tidyestos datos antes de la limpieza de datos (como nombres de limpieza) o para prepararnos para el análisis y la visualización. Para obtener más información sobre el concepto de datos ordenados, lea el artículo Datos de Tidy de Hadley Wickham . También se recomienda encarecidamente el nuevo libro R for Data Science de Garrett Grolemund y Hadley Wickham (ver Capítulo 12).

Para ordenar los datos de patentes, normalmente tendremos que hacer dos cosas.

1. Separe los datos para que cada celda contenga un punto de datos único (por ejemplo, un nombre, código o número de publicación). Esto normalmente implica separar los datos en columnas.
2. Recopilación de los datos nuevamente. Esto implica transformar los datos en las columnas que hemos creado en filas.

Separar datos en columnas es muy fácil en herramientas como Excel. Sin embargo, reunir los datos de nuevo en filas separadas es muy difícil. Afortunadamente, esto es muy fácil de hacer en R con el tidyrpaquete.

El tidyrpaquete contiene tres funciones que son muy útiles cuando se trabaja con datos de patentes. Cuando se trata de campos concatenados en columnas, la función clave es separate_rows.

Aquí trabajaremos con el applicants_cleanedcampo en el conjunto de datos de pizza. Este campo contiene nombres concatenados con a ;como separador. Por ejemplo, en las líneas 1_9 hay nombres de un único solicitante o valores de NA. Sin embargo, en las líneas 10 y 59 vemos:

```
Ehrno Flexible A/S; Stergaard, Ole
Farrell Brian; Mcnulty John; Vishoot Lisa
```

El problema aquí es que cuando tratamos con miles de líneas de nombres de solicitantes, no sabemos cuántos nombres pueden concatenarse en cada celda como base para separar los datos en columnas. Una vez que hubiéramos dividido las columnas (por ejemplo, usando Texto a columnas en Excel), tendríamos que resolver cómo reunir las columnas en filas. La separate_rows()función de tidyrhace que la luz funcione de este problema. Para usar la función nombramos el conjunto de datos, la columna que queremos separar en filas y el separador (sep).

```
library(dplyr)
  library(tidyr)
pizza1 <- separate_rows(pizza_1990_2012, applicants_cleaned, sep = ";")
pizza1
## # A tibble: 12,729 × 31
  ##    applicants_cleaned_type   applicants_organisations
    ##                      <chr>                      <chr>
      ## 1 People                          <NA>
     ## 2 Corporate    Lazarillo De Tormes S L
     ## 3 People                         <NA>
     ## 4 People                         <NA>
```

```
     ## 5 Corporate           Frisco Findus Ag
     ## 6 Corporate Bicycle Tools Incorporated
     ## 7 People                       <NA>
    ## 8 People                       <NA>
    ## 9 People                       <NA>
 ## 10 Corporate; People         Ehrno Flexible A/S
 ## # ... with 12,719 more rows, and 29 more variables:
 ## #   applicants_original <chr>, inventors_cleaned <chr>,
 ## #   inventors_original <chr>, ipc_class <chr>, ipc_codes <chr>,
 ## #   ipc_names <chr>, ipc_original <chr>, ipc_subclass_codes <chr>,
 ## #   ipc_subclass_detail <chr>, ipc_subclass_names <chr>,
 ## #   priority_country_code <chr>, priority_country_code_names <chr>,
 ## #   priority_data_original <chr>, priority_date <chr>,
  ## #   publication_country_code <chr>, publication_country_name <chr>,
 ## #   publication_date <chr>, publication_date_original <chr>,
 ## #   publication_day <int>, publication_month <int>,
 ## #  publication_number <chr>, publication_number_espacenet_links <chr>,
 ## #   pubyear <int>, title_cleaned <chr>, title_nlp_cleaned <chr>,
 ## #   title_nlp_multiword_phrases <chr>, title_nlp_raw <chr>,
 ## #   title_original <chr>,
  applicants_cleaned <chr>
```

Nuestro conjunto de datos original contenía 8.262 filas. Nuestro nuevo conjunto de datos dividido en nombres de solicitantes contiene 12.729 filas. La función ha movido nuestra columna de destino de la columna 1 a la columna 31 en el marco de datos. Podemos moverlo de vuelta fácilmente para inspeccionarlo.

```
library(dplyr)  pizza1 <- select(pizza1, 31, 1:30)
```

separate_rows()ha hecho un gran trabajo, pero uno de los problemas con los nombres concatenados es el espacio en blanco adicional alrededor del separador. Nos ocuparemos de esto a continuación.

12.9.1 Recorte constringr

Si inspeccionamos la parte inferior de la columna mediante su subconjunto $, veremos que muchos de los nombres tienen un espacio inicial en blanco. Esto resulta del ejercicio separado donde ;está realmente ;space. Echa un vistazo a las últimas filas de los datos utilizando tail().

```
tail(pizza1$applicants_cleaned, 20)
##  [1] "Yahoo! Inc"
  ##  [2] "Clarcor Inc"
  ##  [3] "Holden Jeffrey A"
  ##  [4] " Vengroff Darren E"
  ##  [5] "Casper Jeffrey L"
  ##  [6] " Erickson Braden J"
  ##  [7] " Oppenheimer Alan A"
  ##  [8] " Ray Madonna M"
  ##  [9] " Weber Jean L"
   ## [10] "Pandey Neena"
   ## [11] " Sharma Sudhanshu"
   ## [12] " Verizon Patent And Licensing Inc"
   ## [13] "Pandey Neena"
   ## [14] " Sharma Sudhanshu"
   ## [15] "Brown Michael"
   ## [16] " Urban Scott"
   ## [17] "Brown Michael"
   ## [18] " Urban Scott"
   ## [19] "Cole Lorin R"
   ## [20] " Middleton Scott W"
```

Este es un gran problema porque todos los recuentos que hagamos más adelante al utilizar el campo Applicants_cleaned tratarán a "Oppenheimer Alan A" y "Oppenheimer Alan A" como nombres separados cuando deban agruparse.

Podemos abordar esto en un par de maneras. Un enfoque es reconocer que, en realidad, nuestro separador no es simple, ";"sino ";space"nuestro llamado a separate_rows(). En ese caso, la llamada separate_rows()sería en realidad sep = "; ". Agregaremos una línea de código para ilustrar el impacto de este cambio.

```
tmp <- separate_rows(pizza_1990_2012, applicants_cleaned, sep = "; ")
tail(tmp$applicants_cleaned, 20)
##  [1] "Yahoo! Inc" "Clarcor Inc"
  ##  [3] "Holden Jeffrey A" "Vengroff Darren E"
  ##  [5] "Casper Jeffrey L" "Erickson Braden J"
  ##  [7] "Oppenheimer Alan A" "Ray Madonna M"
  ##  [9] "Weber Jean L" "Pandey Neena"
  ## [11] "Sharma Sudhanshu" "Verizon Patent And Licensing Inc"
```

```
  ## [13] "Pandey Neena" "Sharma Sudhanshu"
  ## [15] "Brown Michael" "Urban Scott"
  ## [17] "Brown Michael" "Urban Scott"
  ## [19] "Cole Lorin R" "Middleton Scott W"
```

Otra forma de abordar esto, es usar la str_trim()función del stringrpaquete.

Podemos solucionar este problema utilizando una función del stringrpaquete str_trim(). Tenemos una opción con respecto str_trim()a si recortar el espacio en blanco a la derecha, a la izquierda o ambos. Aquí elegiremos ambos.

Debido a que estamos buscando modificar una columna existente (no para crear un nuevo vector o data.frame) usaremos $para seleccionar la columna y como los datos para la str_trim()función. Eso aplicará la función a la columna de solicitantes en pizza1.

```
library(stringr)
pizza1$applicants_cleaned <- str_trim
(pizza1$applicants_cleaned, side = "both")
  tail(pizza1$applicants_cleaned, 20)
## [1] "Yahoo! Inc" "Clarcor Inc"
  ## [3] "Holden Jeffrey A" "Vengroff Darren E"
  ## [5] "Casper Jeffrey L" "Erickson Braden J"
  ## [7] "Oppenheimer Alan A" "Ray Madonna M"
  ## [9] "Weber Jean L" "Pandey Neena"
  ## [11] "Sharma Sudhanshu" "Verizon Patent And Licensing Inc"
  ## [13] "Pandey Neena" "Sharma Sudhanshu"
  ## [15] "Brown Michael" "Urban Scott"
  ## [17] "Brown Michael" "Urban Scott"
  ## [19] "Cole Lorin R" "Middleton Scott W"
```

Tenga en cuenta que cuando str_trim()usamos usamos subconjuntos para modificar la columna de solicitantes en su lugar. Posiblemente haya una forma más eficiente de hacer esto con tuberías, pero esto parece difícil porque el data.frame debe existir para str_trim()que actúe en su lugar o terminamos con un vector de nombres de solicitantes en lugar de un data.frame. Se proporciona una solución a este problema en Stack Overflow 1 .

En la práctica, la solución más eficiente en este caso es reconocer que el separador para separate_rowses ";space". Sin embargo, eso no siempre será verdad, haciendo que las herramientas sean stringrinvaluables. Para obtener más información sobre

la manipulación de cuerdas en R, consulte el Capítulo 14 de R para Data Science de Garrett Grolemund y Hadley Wickham .

Podemos unir los pasos hasta el momento utilizando tuberías en el siguiente código más simple que nos convertiremos en la tabla de solicitantes para su uso en la infografía. Agregaremos una llamada para cambiar el nombre y cambiar el nombre de solicitantes, limpiado para poner en orden.

```
library(dplyr)
library(tidyr)
library(stringr)
applicants <- rename(pizza, pubyear = publication_year)
%>% filter(pubyear >=
    1990, pubyear <= 2012) %>% separate_rows(applicants_cleaned,
 sep = "; ") %>%
    rename(applicants = applicants_cleaned) %>% select(31, 1:30)
# moves separated column to the beginning
applicants
## # A tibble: 12,729 × 31
 ## applicants applicants_cleaned_type
 ## <chr> <chr>
 ## 1 <NA> People
## 2 Lazarillo De Tormes S L Corporate
## 3 <NA> People
## 4 Depoortere, Thomas People
## 5 Frisco Findus Ag Corporate
## 6  Bicycle Tools Incorporated Corporate
## 7 Castiglioni, Carlo People
## 8 <NA> People
## 9 Bujalski, Wlodzimierz People
## 10 Ehrno Flexible A/S Corporate; People
## # ... with 12,719 more rows, and 29 more variables:
## # applicants_organisations <chr>, applicants_original <chr>,
## # inventors_cleaned <chr>, inventors_original <chr>, ipc_class <chr>,
 ## # <chr>, ipc_names <chr>, ipc_original <chr>,
 ## # ipc_subclass_codes <chr>, ipc_subclass_detail <chr>,
 ## # ipc_subclass_names <chr>, priority_country_code <chr>,
 ## #   priority_country_code_names <chr>, priority_data_original <chr>,
```

```
## # priority_date <chr>, publication_country_code <chr>,
## # publication_country_name <chr>, publication_date <chr>,
## #   publication_date_original <chr>, publication_day <int>,
## # publication_month <int>, publication_number <chr>,
## # publication_number_espacenet_links <chr>, pubyear <int>,
## # title_cleaned <chr>, title_nlp_cleaned <chr>,
## # title_nlp_multiword_phrases <chr>, title_nlp_raw <chr>,
## # title_original <chr>
```

Queremos crear un gráfico con los datos de los solicitantes en nuestro software de infografía. Para eso necesitamos introducir un campo con el que contar. Es posible que también queramos establecer un punto de corte en función del número de registros por solicitante.

En este código simplemente imprimiremos los solicitantes clasificados en orden descendente. La segunda a la última línea del código proporciona un filtro en el número de registros. Este valor se puede cambiar después de inspeccionar los datos. La línea final omite los valores de NA (de lo contrario, el resultado superior) donde el nombre del solicitante no está disponible.

```
library(tidyr)
     library(dplyr)
     applicant_count
 <- select(applicants, applicants, publication_number)
 %>% mutate(n = sum(publication_number = 1)) %>%
    count(applicants, wt = n) %>% arrange(desc(nn))
 %>% filter(nn >= 1) %>%  na.om
            it()  applicant_count
## # A tibble: 6,178 × 2
   ## applicants    nn
   ## <chr> <dbl>
   ## 1  Graphic Packaging International, Inc   154
    ## 2             Kraft Foods Holdings, Inc   132
    ## 3                              Google Inc   123
    ## 4                 Microsoft Corporation    88
    ## 5                 The Pillsbury Company    83
    ## 6                    General Mills, Inc    77
    ## 7                                Nestec    77
    ## 8          The Procter & Gamble Company    59
```

```
## 9                        Pizza Hut, Inc    57
## 10                           Yahoo! Inc    54
## # ... with 6,168 more rows
```

Si inspeccionamos el conteo de solicitantes utilizando View(applicant_count)tenemos 6,178 filas. Eso es demasiado para mostrar en una infografía. Entonces, a continuación, filtraremos los datos sobre el valor para los diez primeros (54). Luego, escribiremos los datos en un archivo .csv usando el simple write_csv()desde readr.

```
library(dplyr)
library(tidyr)
library(readr)
applicant_count <- select(applicants, applicants, publication_number)
           %>% mutate(n = sum(publication_number = 1)) %>%
    count(applicants, wt = n) %>% arrange(desc(nn))
              %>% filter(nn >= 54) %>%


    na.omit()
applicant_count
## # A tibble: 10 × 2
  ##                                applicants    nn
  ##                                     <chr> <dbl>
  ## 1  Graphic Packaging International, Inc   154
  ## 2              Kraft Foods Holdings, Inc   132
  ## 3                             Google Inc   123
  ## 4                 Microsoft Corporation    88
  ## 5                 The Pillsbury Company    83
 ## 6                   General Mills, Inc    77
  ## 7                                Nestec    77
  ## 8          The Procter & Gamble Company    59
  ## 9                         Pizza Hut, Inc    57
   ## 10                           Yahoo! Inc    54
write_csv(applicant_count, "pizza_applicants_1990_2012.csv")
```

Cuando inspeccionemos applicant_count, veremos que Graphic Packaging International es el resultado principal con 154 resultados y Google ocupa el tercer lugar con 123 resultados seguidos por Microsoft. Esto podría sugerir que Google y Microsoft están entrando repentinamente en el mercado de las ventas de pizza en

línea o el software para hacer pizzas o, como es más probable, que existen otros usos de la palabra pizza en los datos de patentes que desconocemos.

Como parte de nuestra infografía, desearemos explorar este intrigante resultado con más detalle. Podemos hacer esto creando un subdataset para Google usando filter().

12.10 Selección de solicitantes utilizandofilter()

Como vimos anteriormente, mientras select()funciona con columnas, filter()desde dplyrtrabajos con filas. Aquí filtraremos los datos para seleccionar las filas en la columna de solicitantes que contienen Google Inc. y luego las escribiremos en un archivo .csv para usar en nuestra infografía. Tenga en cuenta el uso del doble ==y las citas en torno a "Google Inc".

```
library(dplyr)  library(readr)
  google <- filter(applicants, applicants == "Google Inc")
  google
## # A tibble: 123 × 31
  ##    applicants applicants_cleaned_type applicants_organisations
  ##         <chr>                   <chr>                     <chr>
  ## 1  Google Inc       Corporate; People               Google Inc
  ## 2  Google Inc               Corporate               Google Inc
  ## 3  Google Inc       Corporate; People               Google Inc

  ## 4  Google Inc       Corporate; People               Google Inc
  ## 5  Google Inc               Corporate               Google Inc
  ## 6  Google Inc               Corporate               Google Inc
  ## 7  Google Inc               Corporate               Google Inc
  ## 8  Google Inc       Corporate; People               Google Inc
  ## 9  Google Inc               Corporate               Google Inc
  ## 10 Google Inc               Corporate               Google Inc
  ## # ... with 113 more rows, and 28 more variables:
  ## #   applicants_original <chr>, inventors_cleaned <chr>,
  ## #   inventors_original <chr>, ipc_class <chr>, ipc_codes <chr>,
  ## #   ipc_names <chr>, ipc_original <chr>, ipc_subclass_codes <chr>,
  ## #   ipc_subclass_detail <chr>, ipc_subclass_names <chr>,
  ## #   priority_country_code <chr>, priority_country_code_names <chr>,
  ## #   priority_data_original <chr>, priority_date <chr>,
```

```
  ## #   publication_country_code <chr>, publication_country_name <chr>,
  ## #   publication_date <chr>, publication_date_original <chr>,
  ## #   publication_day <int>, publication_month <int>,
  ## #publication_number <chr>, publication_number_espacenet_links <chr>,
  ## #   pubyear <int>, title_cleaned <chr>, title_nlp_cleaned <chr>,
  ## #   title_nlp_multiword_phrases <chr>, title_nlp_raw <chr>,
  ## #   title_original <chr>
write_csv(google, "google_1990_2012.csv")
```

Tenga en cuenta que el resultado correcto para el período 1990 a 2012 para Google es 123 registros de 191 registros en todo el conjunto de datos de pizza. El resultado correcto se logrará solo cuando use los datos filtrados, separados y recortados que creamos en el marco de datos del solicitante.

12.11 Generando tablas IPC

En el siguiente paso, querremos generar dos tablas que contengan datos de la Clasificación Internacional de Patentes (IPC). Los códigos IPC y la Clasificación de Patentes Cooperativas (CPC, no presente en este conjunto de datos) proporcionan información sobre las tecnologías involucradas en un documento de patente. El IPC es jerárquico y procede del nivel de clase general al nivel de grupo y subgrupo detallado. La experiencia revela que la mayoría de los documentos de patentes reciben más de un código IPC para describir con más detalle los aspectos tecnológicos de los documentos de patentes.

El conjunto de datos de pizza contiene códigos IPC en la clase y el nivel de subclase en campos concatenados. Una consideración importante al usar los datos de IPC es que las descripciones son largas y pueden ser difíciles de entender para los no especialistas. Esto puede dificultar la visualización de los datos y, a menudo, requiere esfuerzos manuales para editar las etiquetas para su visualización.

Ahora queremos generar tres tablas IPC.

1. Una tabla general de IPC para el conjunto de datos de pizza
2. Una tabla general de IPC para el conjunto de datos de Google
3. Una tabla de subclases de IPC más detallada para el conjunto de datos de Google

Para facilitar la presentación en una infografía utilizaremos el ipc_classcampo. Para muchos propósitos de análisis de patentes, esto será demasiado general. Sin embargo, tiene la ventaja de ser fácil de visualizar.

Para generar la tabla podemos usar una función genérica basada en el código desarrollado para tratar con los datos de los solicitantes. Llamaremos a la función patent_count ().

```
patent_count
 <- function(data, col = "", count_col = "", n_results = n_results,
       sep = "[^[:alnum:]]+") {
     p_count <- dplyr::select_(data, col, count_col) %>%
 tidyr::separate_rows_(col,
           sep = sep) %>% dplyr::mutate_(n = sum(count_col = 1)) %>%
dplyr::select(2:3)
      p_count %>% dplyr::group_by_(col) %>% dplyr::tally() %>%
 dplyr::arrange(desc(nn)) %>%
           dplyr::rename(records = nn) %>% dplyr::ungroup() %>%
 na.omit() %>% .[1:n_results,           ]  }
```

La patent_count()función se basa en el código que desarrollamos para los solicitantes. Contiene variaciones para que funcione como una función. La función toma cuatro argumentos:

1. col = la columna concatenada que queremos dividir y reunir de nuevo en
2. col_count = una columna para generar recuentos (en este conjunto de datos, el número de publicación)
3. n_results = el número de resultados que queremos ver en la nueva tabla (generalmente 10 o 20 para visualización). Esto es equivalente al número de filas que desea ver.
4. sep = el separador a usar para separar los datos en col. Con los datos de patentes, esto es casi siempre ";" (como ;space.

Para generar los ipc_classdatos podemos hacer lo siguiente y luego escribir el archivo en .csv. Tenga en cuenta que hemos establecido el número de resultados n_resultsa 10.

```
pizza_ipc_class <- patent_count(data = pizza_1990_2012,
 col = "ipc_class", count_col = "publication_number",
       n_results = 10, sep = ";")
  pizza_ipc_class
## # A tibble: 10 × 2
 ## ipc_class records
 ## <chr>   <dbl>
```

```
 ## 1 A21: Baking    2218
 ## 2 G06: Computing    1209
 ## 3 A23: Foods Or Foodstuffs    1058
 ## 4 B65: Conveying     903
 ## 5 A23: Foods Or Foodstuffs     785
 ## 6 A47: Furniture     645
 ## 7 B65: Conveying     480
 ## 8 H05: Electric Techniques Not Otherwise Provided For 456
 ## 9 H04: Electric Communication Technique 427
 ## 10 H04: Electric Communication Technique 320
write_csv(pizza_ipc_class, "pizza_ipcclass_1990_2012.csv")
```

Tenga en cuenta que este conjunto de datos se basa en el pizza_1990_2012conjunto de datos principal (incluidos los casos en los que no hay disponible un nombre de solicitante). La razón por la que no hemos utilizado el conjunto de datos de los solicitantes es porque ese conjunto de datos duplicará el campo de IPC para cada división del nombre de un solicitante. Como resultado, contará en exceso los IPC por el número de solicitantes en un nombre de documento. Como esto sugiere, es importante tener cuidado al trabajar con datos que se han ordenado debido al impacto en otros aspectos.

Este problema no se aplica en el caso de nuestros datos de Google porque el único solicitante que figura en esos datos es Google (excluyendo a los solicitantes). Por lo tanto, podemos usar de forma segura el conjunto de datos de Google para identificar los códigos IPC.

```
google_ipc_class
 <- patent_count(data = google,
 col = "ipc_class", count_col = "publication_number",
       n_results = 10, sep = ";")
  google_ipc_class
## # A tibble: 10 × 2
  ##ipc_class records
##<chr>   <dbl>
  ## 1 G06: Computing      95
  ## 2                          G01: Measuring      14
  ## 3                          G09: Educating      11
  ## 4                          G06: Computing      10
  ## 5   H04: Electric Communication Technique      10
  ## 6   H04: Electric Communication Technique       7
```

```
  ## 7                G10: Musical Instruments       6
  ## 8                         G08: Signalling       1
  ## 9                G10: Musical Instruments       1
  ## 10                            A63: Sports       1
write_csv(google_ipc_class, "google_ipcclass_1990_2012.csv")
```

Solo hay 7 clases y, como es de esperar, están dominadas por la informática. Es posible que deseamos profundizar en esto con un poco más de detalle, por lo que también creamos un campo de subclase de IPC.

```
google_ipc_subclass
 <- patent_count(data = google, col = "ipc_subclass_detail",

      count_col = "publication_number", n_results = 10, sep = ";")

google_ipc_subclass
## # A tibble: 10 × 2
    ## ipc_subclass_detail
    ## <chr>
    ## 1 G06F: Electric Digital Data Processing
    ## 2 G01C: Measuring Distances, Levels Or Bearings
    ## 3 G06Q: Data Processing Systems Or Methods,
     Specially Adapted For Administra
    ## 4  G06Q: Data Processing Systems Or Methods,
    Specially Adapted For Administrat
    ## 5 G09B: Educational Or Demonstration Appliances
    ## 6 G06F: Electric Digital Data Processing
    ## 7 H04W: Wireless Communication Networks
    ## 8 G10L: Speech Analysis Or Synthesis
    ## 9 G09G: Arrangements Or Circuits For Control Of
     Indicating Devices Using Sta
    ## 10 H04B: Transmission
     ## # ... with 1 more variables: records <dbl>
write_csv(google_ipc_subclass, "google_ipcsubclass_1990_2012.csv")
```

Ahora tenemos los datos sobre áreas de tecnología que necesitamos para comprender nuestros datos. El siguiente y último paso es generar datos de los campos de texto.

12.11.1 Tablas de frases

Usaremos datos de palabras y frases en los títulos de documentos de patente para usar en una nube de palabras en nuestra infografía. Es posible generar este tipo de datos en R directamente usando los paquetes tmy NLP. Nuestro conjunto de datos de pizza ya contiene un campo de título dividido en frases con el software Vantagepoint y por eso lo usaremos. Usaremos el campo title_nlp_multiword_phrasesya que las frases son generalmente más informativas que las palabras individuales. Una vez más, usaremos nuestra patent_count()función general , aunque es posible que se necesite experimentación para identificar el número de frases que se visualizan bien en una nube de palabras.

```
pizza_phrases
 <- patent_count(data = pizza_1990_2012,
col = "title_nlp_multiword_phrases",
       count_col = "publication_number", n_results = 15, sep = ";")
  pizza_phrases
## # A tibble: 15 × 2
      ##    title_nlp_multiword_phrases records
      ## <chr>   <dbl>
      ## 1                 Food Product     135
      ## 2              Microwave Ovens      99
      ## 3                 Food Product      44
      ## 4                  Crust Pizza      41
       ## 5                conveyor Oven      40
       ## 6              Microwave Ovens      38
       ## 7               Bakery Product      34
       ## 8                  Making Same      33
       ## 9                Baked Product      33
       ## 10                   Cook Food      32
       ## 11                  Pizza Oven      30
       ## 12                   pizza Box      30
       ## 13              Related Method      29
       ## 14           Microwave Cooking      28
       ## 15           microwave Heating      27
write_csv(pizza_phrases, "pizza_phrases_1990_2012.csv")
```

Ahora hacemos lo mismo con los datos de Google.

```
google_phrases
```

```
<- patent_count(data = google, col = "title_nlp_multiword_phrases",
       count_col = "publication_number", n_results = 15, sep = ";")
  google_phrases
## # A tibble: 15 × 2
## title_nlp_multiword_phrases records
## <chr>   <dbl>
## 1 Digital Map System      10
## 2  conversion Path Performance Measures       9
## 3                         Search Results       7
## 4                          Mobile Device       6
## 5                    Location Prominence       4
## 6                     Processing Queries       4
## 7                 Geographical Relevance       4
## 8                   Local Search Results       4
## 9            Network Speech Recognizers       4
## 10                          Search Query       4
## 11                   indexing Documents       3
## 12        providing Profile Information       3
## 13          search Query Categorization       3
## 14                         Search Ranking       3
## 15 aspect-Based Sentiment Summarization       3
write_csv(google_phrases, "google_phrases_1990_2012.csv")
```

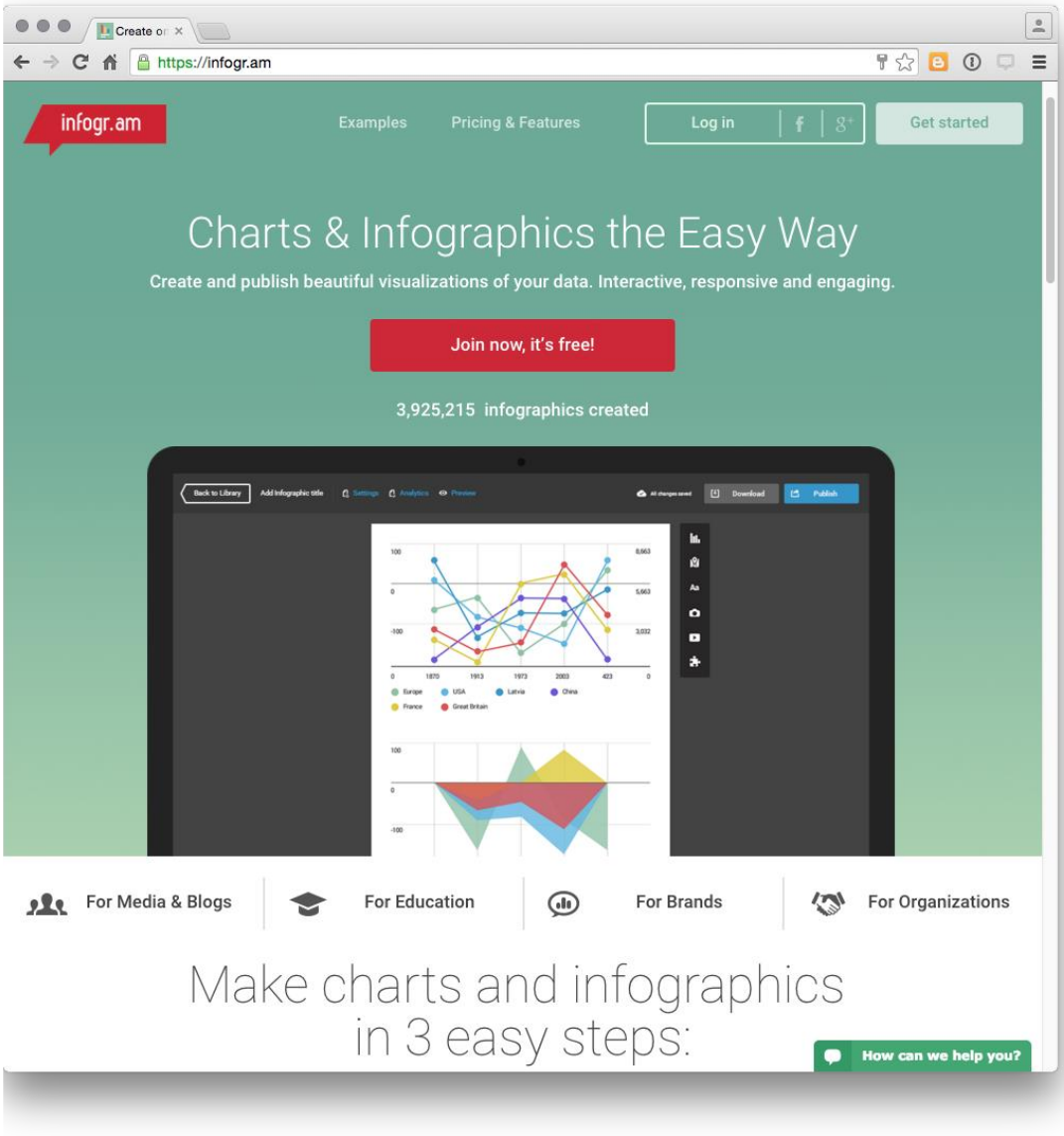
Ahora tenemos los siguientes archivos .csv.

1. pizza_total_1990_2012
2. pizza_country_1990_2012
3. pizza_applicants_1990_2012
4. pizza_ipcclass_1990_2012
5. pizza_phrases_1990_2012
6. Google_1990_2012
7. Google_ipclass_1990_2012
8. Google_ipcsubclass_1990_2012
9. Google_phrases-1990_2012

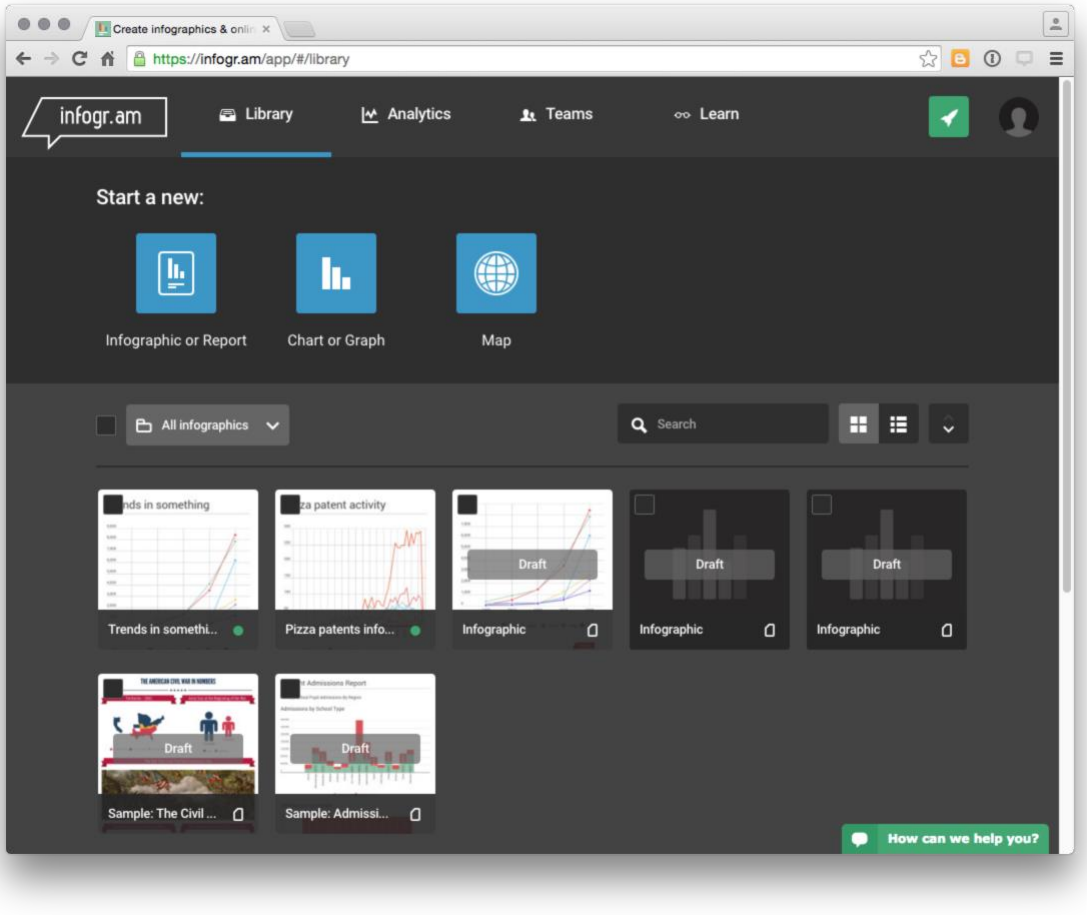
12.12 Creando una infografía en infogr.am

Si está iniciando este capítulo aquí, descargue los conjuntos de datos que usaremos como un solo archivo zip desde el repositorio de manuales aquí y luego descomprima el archivo.

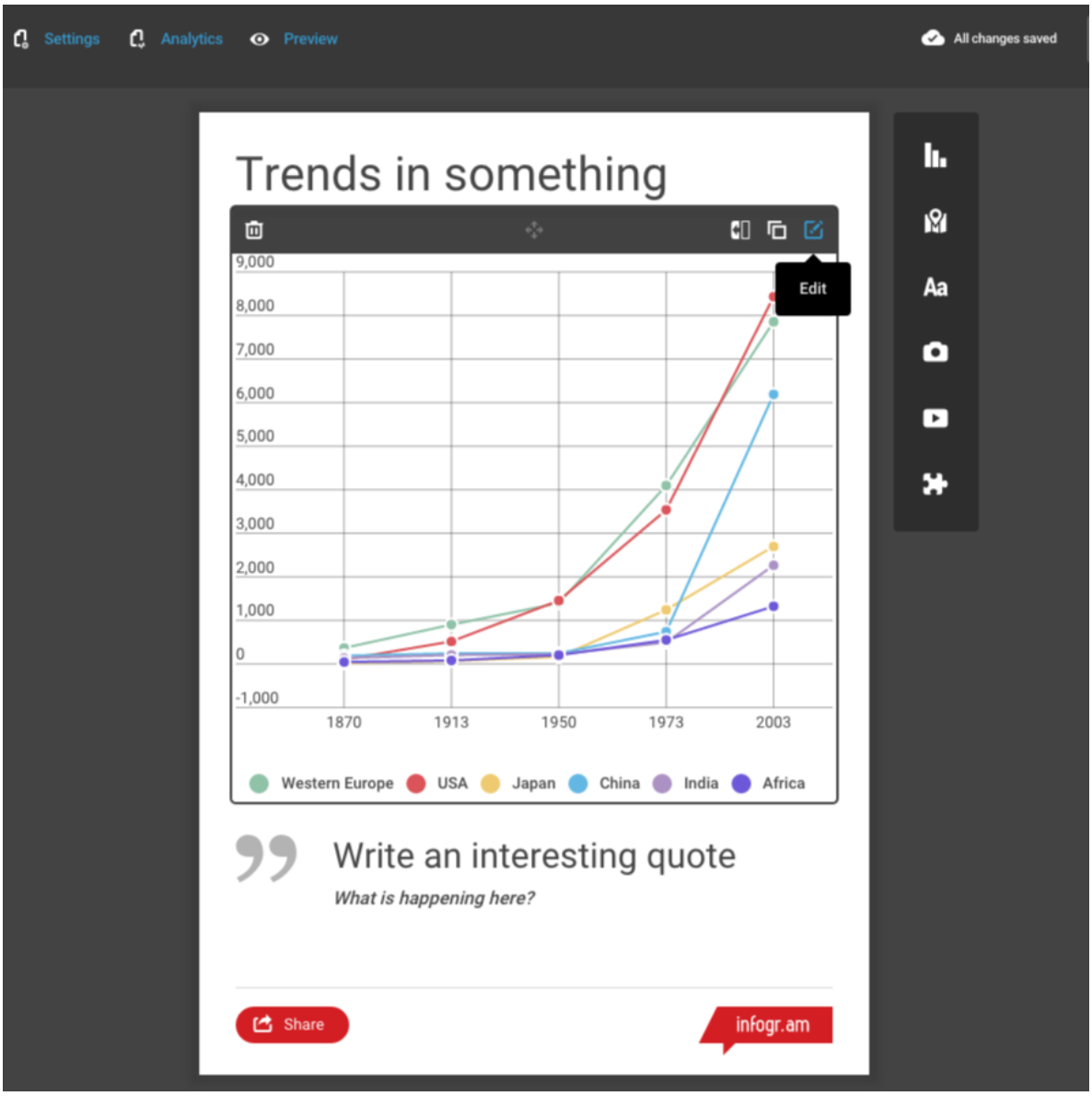
Primero necesitamos registrarnos para obtener una cuenta gratuita con infogr.am



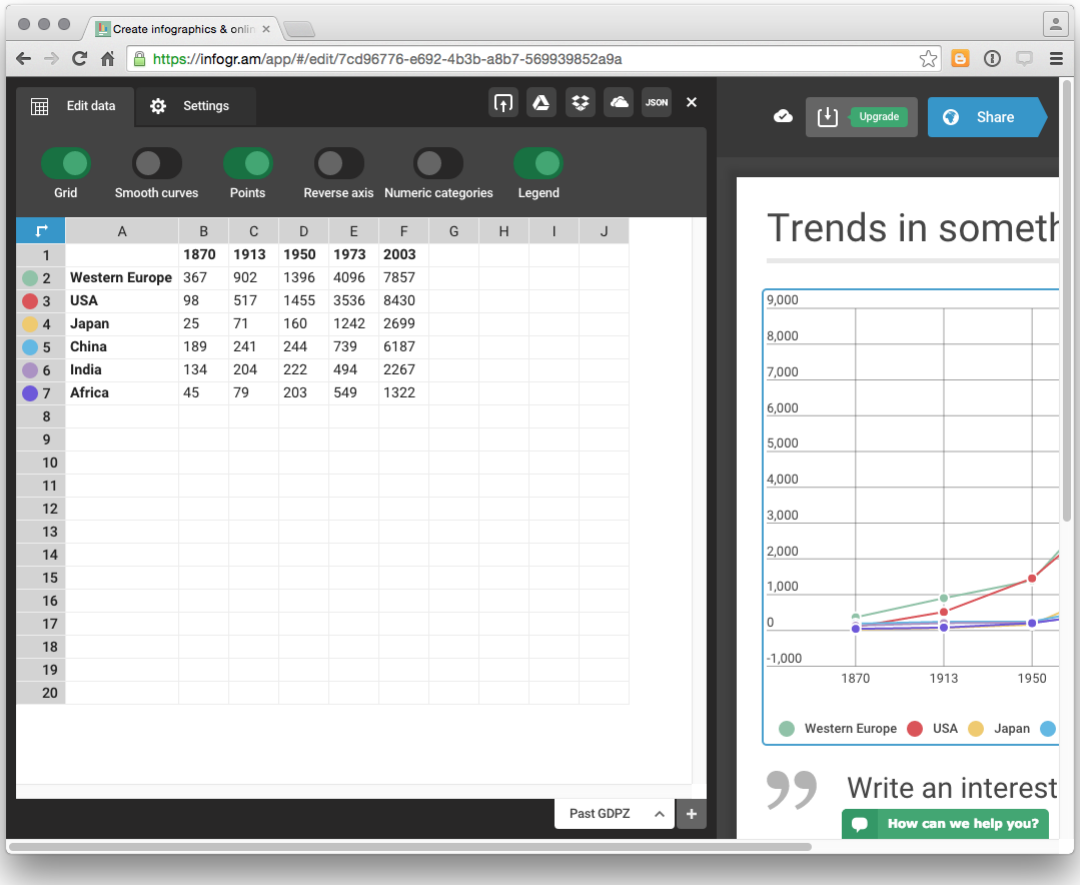
Luego veremos una página con algunos ejemplos de infografías para proporcionar ideas para que pueda comenzar.



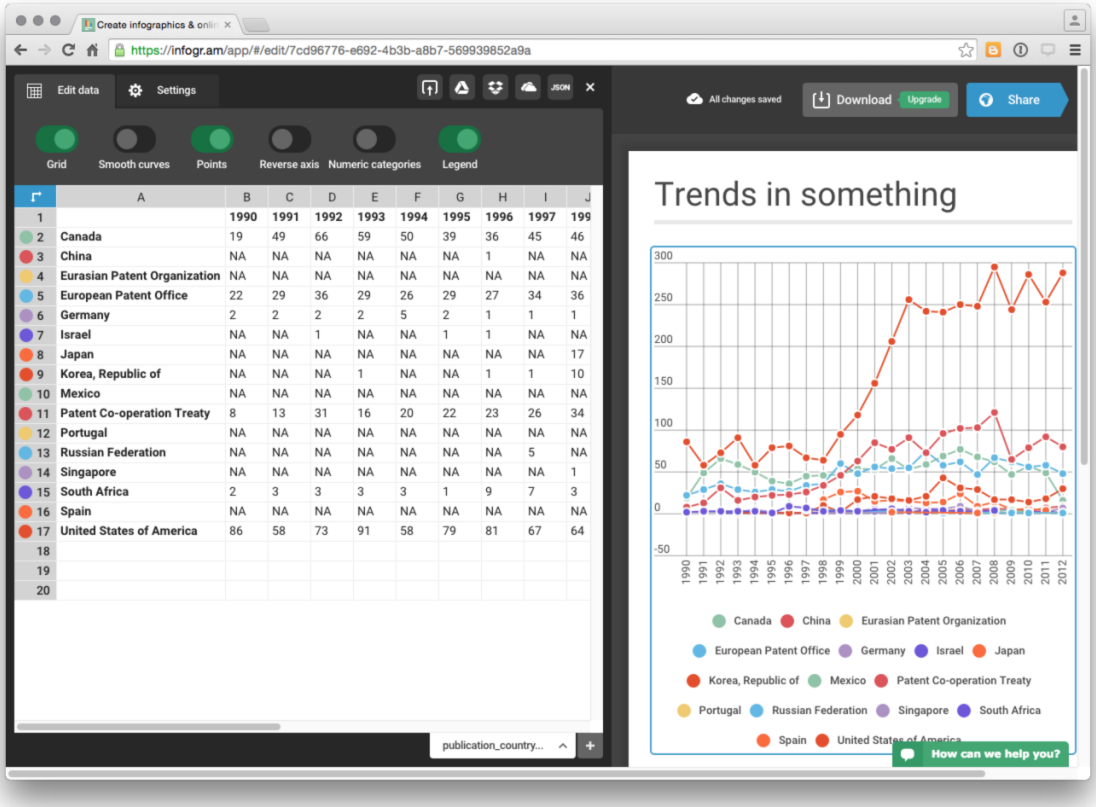
Haga clic en uno de los infogramas con un gráfico como Tendencias en algo y luego haga clic dentro del cuadro del gráfico y seleccione el botón de edición en la parte superior derecha.



Esto abrirá un panel de datos con los datos del juguete visualizados.



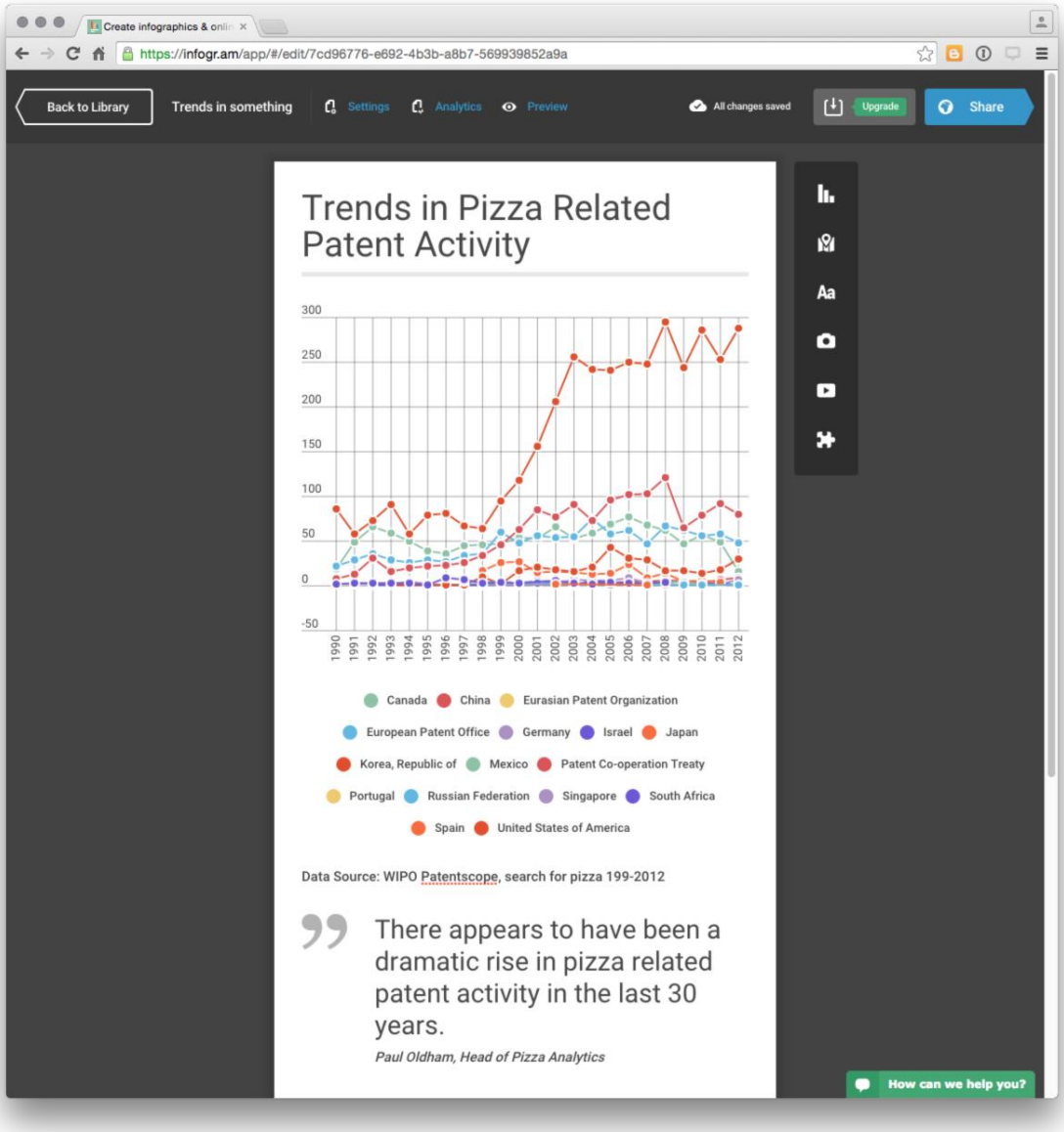
Queremos reemplazar estos datos seleccionando el botón de carga y seleccionando nuestro pizza_country_1990_2012.csvarchivo.



		1990	1991	1992	1993	1994	1995	1996	1997	199
2	Canada	19	49	66	59	50	39	36	45	46
3	China	NA	NA	NA	NA	NA	NA	1	NA	NA
4	Eurasian Patent Organization	NA	NA	NA	NA	NA	NA	NA	NA	NA
5	European Patent Office	22	29	36	29	26	29	27	34	36
6	Germany	2	2	2	2	5	2	1	1	1
7	Israel	NA	NA	1	NA	NA	1	1	NA	NA
8	Japan	NA	NA	NA	NA	NA	NA	NA	NA	17
9	Korea, Republic of	NA	NA	NA	1	NA	NA	1	1	10
10	Mexico	NA	NA	NA	NA	NA	NA	NA	NA	NA
11	Patent Co-operation Treaty	8	13	31	16	20	22	23	26	34
12	Portugal	NA	NA	NA	NA	NA	NA	NA	NA	NA
13	Russian Federation	NA	NA	NA	NA	NA	NA	NA	5	NA
14	Singapore	NA	NA	NA	NA	NA	NA	NA	NA	1
15	South Africa	2	3	3	3	3	1	9	7	3
16	Spain	NA	NA	NA	NA	NA	NA	NA	NA	NA
17	United States of America	86	58	73	91	58	79	81	67	64
18										
19										
20										

Ahora tenemos un gráfico de aspecto decente para los datos de tendencias de nuestros países donde podemos ver la cantidad de registros por país y año al pasar por encima de los puntos de datos relevantes. Si bien algunos de los países con datos de baja frecuencia se encuentran en la parte inferior (y se mostrarían mejor en un gráfico separado), al pasar el mouse sobre los datos o sobre el nombre de un país se mostrará la actividad relevante del país. Por lo tanto, viviremos con esto.

Ahora queremos comenzar a agregar elementos de la historia haciendo clic en el botón de edición en el título. A continuación, podemos comenzar a agregar nuevos cuadros con los iconos de menú de la derecha. Aquí cambiamos el título, agregamos un cuerpo simple para el crédito de datos y luego una cita de alguien que se describe a sí mismo como Jefe de Pizza Analytics.



A continuación, debemos comenzar a profundizar en los datos utilizando nuestros datos de IPC, solicitantes y frases.

Para trabajar con nuestros datos de clase de IPC, agregaremos un gráfico de barras y cargaremos los datos. Para hacer esto, seleccione el ícono del gráfico a la derecha y luego Barra. Una vez más, elegiremos editar y luego cargaremos nuestro pizza_ipcclass_1990_2012conjunto de datos. Luego podemos agregar un cuadro de texto descriptivo. Entonces podemos continuar agregando elementos de la siguiente manera:

1. gráfico de barras de los solicitantes
2. frases de pizza seleccionando gráfico y nube de palabras
3. Subclase de ipc de Google
4. Google nube de palabras.

Un enfoque útil para desarrollar una infografía es comenzar agregando las imágenes y luego agregar títulos y cuadros de texto para resaltar los puntos clave. En el infograma, los cuadros de texto nuevos aparecen debajo de los cuadros existentes, pero se pueden cambiar de posición arrastrando y soltando cuadros unos sobre otros.

Una buena característica del infograma es que es fácil compartir la infografía con otros a través de una url, un código de inserción o en facebook o twitter.

Al final de la infografía, es una buena idea proporcionar un enlace donde el lector pueda obtener más información, como el informe completo o los datos subyacentes. En este caso, agregaremos un enlace al libro de Tableau sobre la actividad de patentes de pizza que desarrollamos en un capítulo anterior .

Nuestra infografía final debería ser algo como esto .

12.12.1 Round Up

En este capítulo nos hemos concentrado en usar R para ordenar los datos de patentes a fin de crear una infografía en línea usando software libre. Usando nuestros datos de patentes de pizza confiables de WIPO Patentscope, pasamos por el proceso de ordenar y ordenar los datos de patentes primero usando líneas cortas de código que luego combinamos en una función reutilizable. Como es de esperar que esta introducción a la ordenación de datos en R haya revelado, R y paquetes como dplyr, tidyry stringrproporcionan herramientas muy útiles para trabajar con datos de patentes, son gratuitos y están bien respaldados.

En la parte final del capítulo, utilizamos los datos que habíamos generado en RStudio para crear una infografía utilizando infogr.am que luego compartimos en línea. Infogram es solo uno de los muchos servicios de infografía en línea y vale la pena probar otros servicios como easel.ly para encontrar un servicio que satisfaga sus necesidades.

Como ya sabrán los usuarios habituales de R, ya es posible producir todos estos gráficos (como las nubes de palabras) directamente en R usando herramientas como ggplot2, plotlyy las nubes de palabras usando paquetes como wordcloud. Algunos de estos temas se han cubierto en otros capítulos y para obtener más

información sobre la minería de textos y las nubes de palabras en R, consulte este artículo reciente sobre R-bloggers . Ninguno de los servicios de infografía que vimos parecía ofrecer una API que permitiera una conexión directa con R. También parece haber una brecha en los paquetes de R, donde la infografía puede aparecer en este artículo de 2015 R-bloggers queproporciona una guía sobre cómo crear Una infografía básica.

1. http://stackoverflow.com/questions/25975827/how-to-feed-the-result-of-a-pipe-chain-magrittr-to-an-object ↩

Capítulo 13 Literatura Científica con Rplos.

13.1 Introducción

En este capítulo analizamos el uso del rplospaquete de rOpenScipara acceder a la literatura científica de la Biblioteca Pública de Ciencias utilizando la API de búsqueda de PLOS .

La Biblioteca Pública de Ciencias (PLOS) es el principal defensor de las publicaciones científicas revisadas por pares de acceso abierto y ha publicado en algún lugar en la región de 140,000 artículos. Estos artículos son un recurso fantástico. PLOS incluye los siguientes títulos.

- MÁS UNO
- Biología del PLOS
- PLOS Medicina
- PLOS biología computacional
- PLOS Genética
- PLOS Patógenos
- PLOS Enfermedades tropicales desatendidas
- Ensayos clínicos PLOS ()
- Colecciones PLOS (colecciones de artículos)

PLOS es importante porque proporciona acceso abierto al texto completo de la investigación revisada por pares. Para los investigadores interesados en trabajar con R rplosy su paquete hermano más grande, el paquete rOpenScifulltext es una herramienta muy importante para acceder a la investigación.

Este artículo es parte del trabajo en curso para el Manual de la OMPI sobre análisis de patentes de código abierto. El objetivo del Manual es introducir herramientas analíticas de código abierto para los investigadores de patentes en países en desarrollo y ser de mayor uso para la comunidad de investigación en ciencia y tecnología. Una parte importante de la investigación de patentes es poder acceder y analizar la literatura científica.

Este artículo no hace suposiciones sobre el conocimiento de R o la programación. rploses un buen lugar para comenzar a aprender cómo acceder a la literatura científica en R usando las interfaces de programación de aplicaciones (API). Debido a que rplosestá bien organizado y los datos están muy limpios, también es

un buen lugar para aprender algunos de los conceptos básicos del trabajo con datos en R. Esto proporciona una buena base para trabajar con el paquete de texto completo de ROpenSci . fulltextle permite recuperar literatura científica de múltiples fuentes de datos y trataremos con eso a continuación.

También vamos a utilizar esto como una oportunidad para introducir algunos de los paquetes populares para trabajar con datos en I, en particular la familia de paquetes para poner en orden y disputas de datos desarrollados por Hadley Wickham en rstudio (es decir, plyr, dplyr, stringry tidyr). Solo los abordaremos, pero incluimos como paquetes de trabajo diarios que le serán útiles para aprender más sobre R.

El primer paso es asegurarse de que tiene R y RStudio.

13.2 Instalar R y RStudio

Para empezar a funcionar, necesita instalar una versión de R para su sistema operativo. Puedes hacerlo desde aquí . Luego descargue RStudio Desktop para su sistema operativo desde aquí usando el instalador para su sistema. Entonces abre RStudio.

13.3 Crear un proyecto

Los proyectos son probablemente la mejor manera de organizar tu trabajo en RStudio. Para crear un nuevo proyecto, seleccione el menú desplegable en la parte superior derecha donde verá el icono azul de R. Desplácese hasta el lugar donde desea guardar sus materiales de R y asigne un nombre a su proyecto (p. Ej., Rplos). Ahora podrá guardar su trabajo en una carpeta de proyecto de rplos y R mantendrá todo junto cuando guarde el proyecto.

13.4 Instalar paquetes

Primero necesitamos instalar algunos paquetes para ayudarnos a trabajar con los datos. Esta lista de paquetes son paquetes "ir a" comunes para uso diario.

```
install.packages("rplos")  #the main event
 install.packages("readr")  #for reading data
 install.packages("plyr")  #for wrangling data
 install.packages("dplyr")  #for wrangling data
 install.packages("tidyr")  #for tidying data
 install.packages("stringr")  #for manipulating strings
```

```
install.packages("tm")  #for text mining
install.packages("XML")  #for dealing with text in xml
```

Luego cargamos las bibliotecas. Tenga en cuenta que rplosinstalará y cargará cualquier otro paquete que necesite (en este caso, ggplot2 para gráficos), por lo que no debemos preocuparnos por eso.

```
library(rplos)


library(plyr)  # load before dplyr to avoid errors
 library(dplyr)
 library(tidyr)
 library(stringr)
library(tm)
 library(XML)
```

A continuación, echemos un vistazo a la amplia gama de funciones disponibles para la búsqueda, rplosmoviéndonos a la pestaña Paquetes en RStudio y haciendo clic en rplos. Un tutorial muy útil sobre el uso rplosse puede encontrar aquí y puede citarse como "Scott Chamberlain, Carl Boettiger y Karthik Ram (2015). rplos: Interfaz para la búsqueda de revistas PLOS. Versión del paquete R 0.5.0 https://github.com/ropensci/rplos ". Si ya te sientes cómodo trabajando en R, deberías dirigirte a ese tutorial introductorio, ya que este artículo contiene muchas más explicaciones. Sin embargo, también agregaremos algunos ejemplos nuevos y código para trabajar con los resultados para agregar a la base de recursos para rplos.

13.5 Funciones clave en rplos.

R es un lenguaje orientado a objetos, lo que significa que funciona en objetos como un vector, tabla, lista o matriz. Estos son fáciles de crear. Luego aplicamos funciones a los datos desde base Ro desde paquetes que hemos instalado para tareas particulares.

- searchplos(), la función básica para la búsqueda de plos.
- plosauthor()buscar el nombre del autor
- plostitle()busca el título
- plosabstract()busca el resumen
- plossubject()buscar por tema
- citations(), busca en el PLOS Rich Citations
- plos_fulltext(), recupera el texto completo usando un DOI

- highplos(), resaltar los términos de búsqueda en los resultados.
- highbrow(), busque términos de búsqueda en un navegador con hipervínculos.

Las funciones en R toman (aceptan) argumentos que son opciones para el tipo de datos que queremos obtener cuando usamos una API o los cálculos que queremos ejecutar en los datos. Para ello rplos, utilizaremos principalmente argumentos para establecer nuestra consulta de búsqueda, los campos que queremos buscar y el número de resultados.

Si eres nuevo en R, esto normalmente toma la forma de un pequeño fragmento de código que está estructurado de esta manera.

```
newobject <- function(yourdata, argument1, argument2, etc)
```

Es probable que un nuevo objeto sea una tabla o lista que contenga datos. el signo <-obtiene o pasa los resultados de la función (como seachplos) al nuevo objeto. Para especificar lo que queremos, primero incluimos nuestros datos (yourdata) y luego uno o más argumentos que controlan lo que obtenemos, como el número de registros o el título, etc.

13.6 Campos de datos en rplos

Hay una gran cantidad de campos que se pueden buscar rploso utilizar para refinar una búsqueda. Sólo usaremos algunos de ellos. Para ver el rango de campos, escriba plosfieldsen la consola y presione Entrar.

```
plosfields
```

Por ejemplo, si quisiéramos buscar el título, el resumen y las conclusiones, usaríamos estos campos para elaborar la consulta (ver más abajo). Si quisiéramos buscar todo menos esos campos, probablemente usaríamos body. Si quisiéramos recuperar las referencias, las incluiríamos referenceen los campos. En rplosun campo se denota fl =con los campos entre comillas como fl = "title"y así sucesivamente, como veremos a continuación.

13.7 Búsqueda básica utilizando searchplos(), navegando y exportando datos

searchplos()es la rplosfunción de búsqueda básica y devuelve una lista de identificadores de documento (DOI) u otros campos de datos. El resultado de la búsqueda básica es un conjunto de DOI que se pueden usar para trabajos

posteriores. Para obtener ayuda para una función, o para encontrar ejemplos de trabajo, use ?delante de la función en la consola:

```
`?`(searchplos)
```

Esto abrirá la página de ayuda para esa función con una descripción de los argumentos que están disponibles y con ejemplos en la parte inferior de la página.

Los ejemplos están ahí para ayudarte. En la rplosactualidad, se centran en el uso de términos de búsqueda única, como la ecología. Sin embargo, como veremos más adelante, es posible usar frases en la búsqueda y usar varios términos. Hay bastantes argumentos (opciones) disponibles para refinar los resultados e incluiremos algunos de estos en los ejemplos.

El autor de este artículo es un gran fanático de la pizza. Entonces, en el primer ejemplo, realizaremos una búsqueda simple del término pizza y luego especificaremos los resultados que queremos ver usando el argumento fl =(para campos) y la cantidad de resultados que queremos ver usando limit = 20. Al especificar los campos utilizaremos c()para combinarlos.

```
p <- searchplos(q = "pizza", fl = c("id",
       "publication_date", "title", "abstract"),       limit = 20)  p
```

Lo que se searchplos()ha hecho en segundo plano es enviar una solicitud a la API de PLOS para recuperar el ID, la fecha de publicación, el título y el resumen de 20 registros en las revistas de PLOS. Para ver el tipo de resultados:

```
p
```

Los resultados en R se almacenan en objetos (en este caso, el objeto es una lista). Para ver el tipo de objeto en R use:

```
class(p)
```

Cuando se trabaja con R, generalmente es más útil comprender la estructura de los datos para que pueda averiguar cómo acceder a ellos. Eso se puede hacer usando str()para la estructura. Esta es una de las funciones más útiles en R y vale la pena escribirla.

```
str(p)
```

Los resultados pueden parecer un poco confusos al principio, pero lo que esto nos dice es que tenemos un objeto R que es una lista que consta de dos componentes. El primero es un elemento llamado metaque informa la cantidad de registros

encontrados y el tipo de objeto (un data.frame). El segundo es el dataque contiene la información sobre los dos resultados en forma de un marco de datos (básicamente una tabla) que contiene la información de identificación, fecha, título y resumen que solicitamos PLOS.

Tenga en cuenta que la lista contiene un marcador $para el comienzo de las dos listas, y los datos que contienen aparecen como ..$indicativos de que están anidadas en metao data. Esta jerarquía nos ayuda a acceder a los datos utilizando subconjuntos en R. Por ejemplo, si quisiéramos acceder a los metadatos (y lo hacemos) podemos usar lo siguiente:

```
p$meta
```

Eso sólo imprimirá las metaentradas de datos completos . Si solo quisiéramos acceder al número de registros (núm $ Encontrado), extenderíamos esto un poco moviéndonos a esa posición en la jerarquía con:

```
p$meta$numFound
```

Eso imprimirá solo el número de registros devueltos por nuestra búsqueda. Una forma alternativa de subconjunto es usar "[" y "[[" y la posición numérica en la lista. En la Programación práctica con RGarrett Grolemund lo compara con un tren con vagones numerados donde "[]" selecciona el vagón del tren, por ejemplo, [1] y "[[1]]" selecciona el contenido del número de carro 1. No necesitamos preocuparse por esto, pero es muy útil como una forma de recordar la diferencia. Por ejemplo, lo siguiente selecciona el contenido del primer elemento en nuestra lista (meta):

```
p[[1]]
```

y es lo mismo que p$meta. Mientras:

```
p[[1]][[1]]
```

es lo mismo que p$meta$numFound.

Subcontratar los datos por su posición numérica en lugar de por su nombre hace que la vida sea mucho más fácil cuando se trabaja con listas con muchos elementos. Como veremos a continuación, cuando apliquemos una función a una lista con múltiples artículos, también podemos usar "[[", 2. Esto recuperará el segundo artículo en cada una de nuestras líneas de vagones de tren.

Otro consejo útil para navegar por los datos en RStudio es usar autocompletar. Intenta escribir lo siguiente en la consola.

Cuando escribimos $ aparece una ventana emergente y se muestran dos entradas como tablas para metay data. Haga clic en meta, luego agregue otro signo $ al final. Ahora mostrará tres elementos en púrpura (para vectores). Selecciona numFoundy hola presto! A medida que trabaje con RStudio, notará que cuando empiece a escribir el nombre de una función, las listas de nombres comenzarán a aparecer. Tiposearchen la consola pero no presione enter y espere un momento. Debería aparecer una lista con tres elementos con búsqueda {base}, rutas de búsqueda {base} y searchplos {rplos}. Esto es realmente útil porque ahorra mucho escribir. A medida que se familiariza con R, también muestra de manera útil lo que hace una función y un recordatorio de sus argumentos. Los soportes blandos alrededor de {base} indican el paquete donde se puede encontrar la función (esto puede ser útil para descubrir funciones cuando te quedas atascado).

Finalmente, también puede ver los elementos de su proyecto en el panel Entorno. Haga clic en la flecha azul para pen el panel Entorno debajo de Valores y verá la estructura de los datos py parte de su contenido.

13.7.1 Creando un nuevo objeto y escribiendo en un archivo

Ok, entonces tenemos una lista con algunos resultados que contienen metay data. Ahora queremos exportar dataa un archivo .csv con el que podemos trabajar en Excel u otro programa.

Si bien querremos anotar el número total de resultados meta, lo que realmente queremos será data. Simplemente podemos crear un nuevo objeto usando el código anterior y asignarlo a un nombre usando <-. Tenga en cuenta que no hay espacio aquí y < -no funcionará.

```
dat <- p$data
```

Si observamos la clase de este objeto (class(dat)) ahora tenemos un data.frame (una tabla) que podemos escribir en un archivo .csv para usar más adelante. Podemos hacer esto fácilmente write.csv()y comenzar por nombrar el objeto que queremos escribir (dat) y luego darle un nombre de archivo. Debido a que creamos un rplosproyecto en RStudio anteriormente (no lo hicimos), el archivo se guardará en la carpeta del proyecto. Si no creó un proyecto o desea verificar el directorio, use:

```
getwd()
```

Esto le mostrará su directorio de trabajo actual. Si no ve el nombre de su rplosproyecto, copie la ruta completa del archivo para que se vea como esto (no olvide "" alrededor de la ruta):

```
setwd("/Users/pauloldham/Desktop/open_source_master/rplos")
```

Ok, ahora sabemos dónde estamos. Entonces, vamos a guardar el archivo.

```
write.csv(dat, "dat.csv", row.names = FALSE)
```

Si abrimos esto en Excel o Open Office Calc, veremos dos entradas en blanco en los campos abstractos. Las celdas en blanco pueden crear problemas de cálculo. Dentro de R podemos manejar esto llenando los espacios en blanco con NA de la siguiente manera [2]. En este caso, nos estamos subdividiendo en dat y luego pedimos a R que identifique las celdas que coinciden exactamente ==con "". Luego llenamos esas celdas en dat con NA (para No disponible).

```
dat[dat == ""] <- NA  dat
```

Entonces podemos simplemente escribir el archivo como antes. Si quisiéramos eliminar las NA que acabamos de presentar, podríamos usarlas write.csv(dat, "dat.csv", row.names = FALSE, na = "")para convertirlas de nuevo en espacios en blanco.

Una forma más rápida de lidiar con la escritura de archivos es usar el readrpaquete reciente , ya que esto no agregará números de fila a los archivos exportados. Aquí vamos a utilizar la write_csv()función.

```
write_csv(dat, "dat.csv")
```

La ventaja de esto readres que es rápido y no requiere el mismo número de argumentos que el estándar write.csv, como especificar nombres de fila o con read.csvcadenas especificadasAsFactors = FALSE.

Finalmente, si quisiéramos escribir la lista completa p, incluido el metaarchivo, podríamos usar:

```
write.csv(p, "p.csv", row.names = FALSE)
```

Ahora hemos recuperado algunos datos que contienen pizza a través de la API PLOS rplosy hemos escrito los datos en un archivo como una tabla que podemos usar más adelante. Ahora pasaremos a algunas cosas más sofisticadas que podemos hacer rplos.

13.8 Límite por diario

Como hemos visto anteriormente, PLOS contiene 7 revistas y en rploslos resultados de una búsqueda puede limitarse a revistas específicas como PLOS ONE o PLOS Biology. Tenga en cuenta que los nombres cortos de las revistas parecen usar el formato antiguo para PLOS que consiste en una combinación de mayúsculas y minúsculas (por ejemplo, PLoSONE no PLOSONE). Una buena manera fácil de encontrar los nombres cortos de revistas es usar:

```
journalnamekey()
```

Aquí limitaremos la búsqueda a PLOS UNO agregando fq =los argumentos y luego el cross_published_journal_keyargumento. Tenga en cuenta que el fq=argumento toma las mismas opciones que fl=. Pero, fq =filtra los resultados devueltos por PLOS a solo aquellos especificados en fq =.

```
pizza <- searchplos(q = "pizza",
 fl = c("id", "publication_date", "title", "abstract"),
 fq = 'cross_published_journal_key:PLoSONE', start = 0, limit = 20)
  head(pizza$data)
```

Hemos recuperado 20 registros aquí usando limit = 20(el valor predeterminado es 10). En general, es una buena idea comenzar con una pequeña cantidad de resultados para probar que estamos obteniendo lo que esperamos en lugar de muchos datos irrelevantes. ¿Y si quisiéramos recuperar todos los resultados? Aquí tendremos que hacer un poco más de trabajo usando el campo meta.

13.9 Obtención del número total de resultados.

Una forma de hacer esto es tomar nuestro número original de resultados y luego subcontratar los datos y crear un nuevo objeto que contenga el valor para el número de registros en numFound. Tenga en cuenta que el número de registros para una consulta en particular a continuación puede haber aumentado en el momento en que lea este artículo.

```
r <- pizza$meta$numFound
```

Para ejecutar una nueva búsqueda, ahora podemos insertar ren el límite = valor. Esto se interpretará como el valor numérico de r(210).

```
pizza <- searchplos(q = "pizza",
 fl = c("id", "publication_date", "title", "abstract"),
```

```
    fq = "cross_published_journal_key:PLoSONE", start = 0, limit = r)
  head(pizza$data)
```

Una forma alternativa de hacer esto es hacernos la vida un poco más fácil ejecutando nuestra consulta y estableciendo el límite como limit = 0. Esto solo devolverá los metadatos. Luego agregamos el subconjunto para el número encontrado al final del código como $meta$numFound. Eso hará retroceder el valor directamente.

```
r <- searchplos(q = "pizza",
 fq = "cross_published_journal_key:PLoSONE", limit = 0)$meta$numFound
  r
```

Luego podemos ejecutar la consulta nuevamente usando el valor de ren limit =:

```
pizza <- searchplos(q = "pizza",
 fl = c("id", "publication_date", "title", "abstract"),
 fq = 'cross_published_journal_key:PLoSONE', start = 0, limit = r)
   head(pizza$data)
```

13.10 Obtención del número de registros en las revistas PLOS

Eso ha devuelto los 210 resultados completos para PLOS ONE. Podríamos intentar hacer la vida aún más fácil obteniendo primero los resultados en todas las revistas de PLOS. Hacemos esto eliminando el fq =argumento que limita los datos a PLOS UNO y guardando el resultado y el objeto al que llamaremos r1. Tenga en cuenta que el número de registros probablemente habrá aumentado para cuando lea esto.

```
r1 <- searchplos("pizza", limit = 0)$meta$numFound  r1
```

Esto produce 298 resultados en el momento de la escritura. ¿Qué sucede ahora si ejecutamos nuestra consulta original utilizando el valor de r1(298 registros) pero limitando los resultados solo a PLOS ONE?

```
pizza <- searchplos(q = "pizza",
 fl = c("id", "publication_date", "title", "abstract"),
 fq = 'cross_published_journal_key:PLoSONE', start = 0, limit = r1)
  pizza$meta$numFound
```

La respuesta es que los 210 resultados en PLOS UNO se devuelven del total de 244 en todas las revistas de PLOS. ¿Por qué? La razón por la que funciona es que

searchplos()inicialmente retira todos los datos de la API de PLOS y luego aplica nuestra entrada fq =como filtro. Entonces, en realidad, los 244 registros completos se recuperan y luego se filtran hasta el 210 desde PLOS ONE. En este caso, esto hace que nuestra vida sea más fácil porque podemos usar los resultados en las publicaciones de PLOS y luego restringir los datos.

13.11 Escribiendo los resultados y usando un libro de códigos

Ahora tenemos un total de 210 resultados para pizza. Simplemente podemos escribir los resultados en un archivo .csv.

```
write.csv(pizza, "plosone_pizza.csv", row.names = FALSE)
```

Como lo ilustra, es muy fácil de usar rplos()y crea rápidamente un archivo que puede usarse para otros fines.

Cuando trabaje en R, a menudo creará varias tablas y dará varios pasos. Para hacer un seguimiento de lo que haces, es una buena idea crear un archivo de texto como un libro de códigos. Usa el libro de códigos para anotar los pasos importantes que das. La idea de un libro de códigos está tomada de Elements of Data Analytic Sytle de Jeffrey Leek, que proporciona una introducción muy accesible para mantenerse organizado. Para crear un libro de códigos en RStudio simplemente use File > New File > Text File. Esto abrirá un archivo de texto que se puede guardar con su proyecto. El libro de códigos le permite recordar qué acciones realizó en los datos meses o años más tarde. También permite que otros sigan y reproduzcan sus resultados y es importante para la investigación reproducible .

13.12 búsqueda de proximidad

Por lo general, desearemos realizar una búsqueda recuperando primero un conjunto de resultados de trabajo aproximado para obtener una idea de los datos y luego experimentar hasta que estemos contentos con la relación datos a ruido (consulte este artículo para ver un ejemplo).

Al pensar en formas de refinar nuestros criterios de búsqueda, también podemos utilizar la búsqueda por proximidad. La búsqueda de proximidad se centra en la distancia entre las palabras que nos interesan. Para leer más sobre este uso ?searchplosen la consola y desplazarse hacia abajo hasta el ejemplo siete en la lista de ayuda. Reproducimos ese ejemplo aquí usando las palabras sintético y biología como nuestros términos.

Podemos establecer la proximidad de los términos utilizando tilde ~y un valor. Por ejemplo, ~15encontrará ejemplos de los términos sintético y biología dentro de 15 palabras el uno del otro en los textos completos de los artículos de PLOS.

```
searchplos(q = "everything:\"synthetic biology\"~15",
 fl = "title", fq = "doc_type:full")
```

Tenga en cuenta que mientras que la síntesis y la biología aparecen entre comillas (lo que sugiere que son una frase para buscar) en realidad, la API tratará esto como una biología sintética. Es decir, la consulta buscará primero los documentos que contengan las palabras biología sintética Y, y luego los casos en que las palabras aparezcan dentro de las 15 palabras una de la otra. En este caso, obtenemos 1,684 resultados a través de PLOS (todo) y textos completos (fq = "doc_type:full) como podemos ver en este código.

```
searchplos(q = "everything:\"synthetic biology\"~15",
 fl = "title", fq = "doc_type:full")$meta$numFound
```

Podemos limitar el horizonte de búsqueda a ~ 1 para capturar aquellos casos en que los términos aparecen uno al lado del otro (dentro de 1 palabra a la izquierda o a la derecha) que produce 1001 resultados.

```
searchplos(q = "everything:\"synthetic biology\"~1",
     fl = "title", fq = "doc_type:full")$meta$numFound
```

En realidad, esto es aproximadamente 10 registros más alto que el total obtenido en una coincidencia exacta para la frase que sugiere que podría haber casos de "biología sintética" u otros problemas (como la puntuación) o el rendimiento de la API que explique la varianza. Como se señala en la searchplos()documentación:

"No se sorprenda si las consultas que realiza en un lenguaje de scripting, como el uso de rplos en R, dan resultados diferentes a los de la búsqueda de artículos en el sitio web de PLOS. No estoy seguro de qué valores predeterminados utilizan exactamente en su sitio web ".

Como resultado, es una buena idea probar diferentes enfoques. Incluso si no es posible llegar al fondo de cualquier variación, es muy útil anotarlo en su libro de códigos para resaltar el problema a otros que pueden intentar repetir su trabajo.

También es importante enfatizar que al usar rplos()es posible devolver un fragmento del texto con los términos resaltados usando highplos()y el hl.fragsizeargumento para establecer el horizonte para el fragmento de texto alrededor de la búsqueda. Esto es particularmente útil para la minería de texto.

En muchos casos, la información más útil proviene de la búsqueda con frases y varios términos. A diferencia de las palabras, las frases pueden articular conceptos. Esto generalmente los hace más útiles que las palabras individuales para buscar información.

13.13 Buscando usando frases múltiples

Para buscar por frases, comenzamos creando un objeto que contiene nuestras frases y ponemos las frases entre comillas dobles. Si no utilizamos comillas dobles, la búsqueda buscará documentos que contengan ambas palabras en lugar de la frase completa (p. Ej., Biología sintética y no "biología sintética"). Tenga en cuenta que el código a continuación mostrará "" como "" pero no necesita ingresar el \.

En este ejemplo , utilizaremos la consulta de búsqueda desarrollada en este artículo de PLOS ONE sobre biología sintética y recuperaremos la identificación, los datos, el autor, el título y el resumen en las publicaciones de PLOS.

Primero creamos la consulta de búsqueda. Tenga en cuenta que usamos c(), para combinar, para combinar la lista de términos en un vector dentro del objeto llamado s.

```
s <- c("\"synthetic biology\"", "\"synthetic genomics\"",
 "\"synthetic genome\"",       "\"synthetic genomes\"")  s
```

Ahora queremos obtener el número máximo de resultados devueltos por uno de los términos de búsqueda. Esto es un poco complicado porque rplosdevolverá una lista que contiene cuatro elementos de lista (uno para cada uno de nuestros términos de búsqueda). Cada una de esas listas contendrá metay dataelementos. Lo que queremos hacer es averiguar cuál de los términos de búsqueda devuelve el mayor número de resultados dentro metade numFound. Entonces podemos usar ese número como nuestro límite.

Esto implica más de un paso.

1. Primero tenemos que buscar los datos.
2. Entonces necesitamos extraer metade cada lista.
3. Luego debemos seleccionar, numFoundbuscar y devolver el valor máximo en todas las listas de resultados.

La forma más fácil de hacerlo es crear una pequeña función a la que llamaremos plos_records. Para cargar la función en su entorno, cópielo y péguelo en su consola y presione enter. Los comentarios que siguen #explican lo que sucede se ignorarán

cuando se ejecute la función. Cuando hayas hecho esto, si te mueves a Entorno, verás plos_recordsen Funciones.

```
plos_records <- function(q) {    library(plyr) #for ldply
    library(dplyr) #for pipes, select and filter
      lapply(q, function(x) searchplos(x, limit = 0)) %>%
      ldply("[[", 1) %>% #get meta from the lists
      select(numFound) %>% #select numFound column of meta
      filter(numFound == max(numFound)) %>% #filter on max numFound
        print() #print max value of numFound  }
```

Ahora podemos ejecutar el siguiente código usando snuestra consulta (q = s) en la función. Si todo va bien, se imprimirá un resultado en la consola con el número máximo de resultados. Los resultados pueden tardar unos minutos en volver desde la API.

```
r2 <- plos_records(q = s)  r2
```

Ahora debería ver un número alrededor de 1151 (en el momento de escribir este documento). ¡Hurra!

Ahora podemos usar r2en el límite para devolver todos los registros. Escribiremos esto de manera estándar y luego mostraremos una forma más simple usando las tuberías a %>%continuación. Tenga en cuenta que utilizamos scomo nuestros términos de búsqueda (ver q = s) y que hemos utilizado r2para el límite (límite = r2). Debido a que estamos llamando a una porción de datos, esto puede demorar alrededor de un minuto en ejecutarse.

Tenga en cuenta que en cada paso del código siguiente estamos creando y luego sobrescribiendo un objeto llamado results. También estamos nombrando resultscomo el primer argumento en cada paso. Esto puede tardar unos instantes en ejecutarse.

```
library(plyr)
  results <- lapply(s, function(x)
 searchplos(x, fl = c('id','author', 'publication_date', 'title',
 'abstract'), limit = r2))
  results <- setNames(results, s)
 #add query terms to the relevant results in the list
  results <- ldply(results, "[[", 2)
 #extract the data into a single data.frame
```

Podemos simplificar la vida utilizando tuberías %>%para simplificar el código. La ventaja de usar tuberías es que no tenemos que seguir creando y sobrescribiendo objetos temporales (ver más arriba results). El código también es mucho más fácil de leer y más rápido. Para obtener más información sobre el uso de tuberías, consulte este artículo de Sean Anderson . De nuevo, la consulta puede ser un poco lenta, ya que los datos se recuperan.

```
library(plyr)  library(dplyr)
  results <- lapply(s, function(x)
searchplos(x, fl =
c('id', 'author', 'publication_date', 'title', 'abstract'), limit = r2))
 %>%      setNames(s) %>%       ldply("[[", 2)  results
```

Las tuberías son una innovación relativamente reciente en R (consulte el magrittr, dplyry tidyrpaquetes) y la mayoría del código que verá se escribirá de la manera tradicional. Sin embargo, las tuberías hacen que el código R sea más rápido y mucho más fácil de seguir. Si bien debe estar familiarizado con el código R regular para seguir la mayoría del trabajo existente, las tuberías se están volviendo cada vez más populares porque el código es más simple y tiene una lógica más clara (p. Ej., Haga esto más o menos).

Ahora tenemos nuestros datos que constan de 1,405 registros en un solo marco de datos que podemos ver.

```
View(results)
```

Ahora podríamos simplemente escribir esto en un archivo .csv. Pero hay una serie de cosas que podríamos querer hacer primero. La mayoría de estas tareas se incluyen en la categoría de manejar y ordenar los datos para que podamos seguir trabajando con ellos en R u otro software como Excel.

13.14 Poner en orden y organizar los datos

Muchos depuración de los datos útiles y tareas de organización se pueden realizar fácilmente usando el dplyr()y tidyr()paquetes desarrollados por Hadley Wickham en rstudio. Otros paquetes importantes incluyen stringr()(para trabajar con cadenas de texto), plyr()y reshape2()(disputas generales) y lubridate()(para trabajar con fechas). Estos paquetes fueron desarrollados por Hadley Wickham y sus colegas con el objetivo específico de facilitar el trabajo con los datos en R de manera consistente. Utilizaremos principalmente dplyry tidyren los ejemplos a continuación, una hoja de trucos de RStudio muy útil puede ayudarlo a trabajar con dplyry tidyr.

13.14.1 Renombrando una columna

En primer lugar, podríamos ordenar el cambio de nombre de una columna. Por ejemplo, podríamos querer cambiar el nombre .ida algo más significativo. Podemos usar rename()desde dplyr()para hacer eso (ver ?rename).

```
results <- rename(results, search_terms = .id)  results
```

13.15 Rellenar espacios en blanco

Es una buena práctica llenar celdas en blanco con NA para "No disponible" "para evitar problemas de cálculo. Por ejemplo, como en el ejemplo anterior, tenemos algunas celdas en blanco en el campo de resumen y puede haber otras en otro lugar. Después de este StackOverflow respuesta podemos hacer esto fácilmente.

```
results[results == ""] <- NA
```

Si por alguna razón quisiéramos eliminar los valores de NA, podemos manejar eso en el momento de exportar a un archivo (ver arriba).

13.16 Fechas de conversión

El publication_datecampo es un vector de caracteres. Podemos convertirlo fácilmente en un formato de fecha que se puede usar en R y soltar el T00: 00: 00 para obtener información de la hora usando:

```
results$publication_date <- as.Date(results$publication_date)
  head(results$publication_date)
```

Añadiendo columnas

Cuando se trata de fechas, es posible que deseamos simplemente dividir el publication_datecampo en tres columnas por año, mes y día. Podemos hacer eso usando separate()desde tidyr.

```
results <- separate(results, publication_date,
c("year", "month", "day"), sep = "-",
      remove = FALSE)  head(select(results, year, month, day))
```

Aquí hemos especificado los datos (resultados), la columna que queremos separar (resultados) y luego las tres nuevas columnas que queremos crear al cerrarlas c()y colocarlas entre comillas. Esto crea tres nuevas columnas. El removeargumento

especifica si queremos eliminar la columna original (el valor predeterminado es VERDADERO) o mantenerla.

Debido a que trabajar con fechas puede ser bastante incómodo (por decirlo suavemente), tiene sentido tener un rango de opciones disponibles al principio para trabajar con sus datos en lugar de tener que volver al principio mucho más tarde.

13.17 Añadir una cuenta

Una característica de retirar la literatura de una API para la literatura científica es que los campos tienden a ser campos de caracteres en lugar de numéricos. Los vectores de caracteres en R se citan con "". Esto puede hacer que la vida sea incómoda si queremos comenzar a contar las cosas más adelante. Para agregar una columna de recuento, podemos usarla mutatedel dplyr()paquete para crear una nueva columna number. numberse basa en asignar el valor 1 a las columnas de ID usando mutate(). Estamos evitando el término conteo porque es el nombre de una función count(). Hay otras formas de hacerlo, pero este enfoque apunta a la mutate()función muy útil dplyrpara agregar una nueva variable.

```
library(dplyr)  results <- mutate(results, number = sum(id = 1))
  head(select(results, title, number))
```

Cuando veamos los resultados, ahora veremos un nuevo número de columna que contiene el valor 1 para cada entrada.

13.18 Eliminar una columna

A menudo terminaremos con más datos de los que deseamos, o crearemos más columnas de las que necesitamos. La forma estándar de eliminar una columna es usar la confianza $para seleccionar la columna y asignarla a NULL.

```
results$columnname <- NULL  #dummy example
```

Otra forma de hacer esto, que se puede usar para varias columnas, es usar select()desde dplyr(ver ?select()). Seleccionar solo mantendrá las columnas que nombramos. Podemos hacer esto usando los nombres de columna o posición. Por ejemplo, lo siguiente mantendrá las primeras 8 columnas (1: 8) pero eliminará la novena columna sin nombre porque el valor predeterminado es eliminar las columnas que no tienen nombre. También podríamos escribir los nombres de las columnas, pero usar los números de posición es más rápido en este caso.

```
test <- select(results, 1:8)  length(test)
```

También podríamos colocar columnas por posición usando lo siguiente (para eliminar las columnas 5 y 6). Este enfoque es útil cuando hay muchas columnas que tratar.

```
test <- select(results, 1:4, 7:9)
```

Un enfoque más sencillo en este caso es eliminar columnas de forma explícita -y conservar las demás.

```
test <- select(results, -month, -day)
```

Seleccionar también es muy útil para reordenar columnas. Imaginemos que queríamos mover la idcolumna a la primera columna. Simplemente podemos colocar idcomo primera entrada select()y luego las columnas totales para reordenar.

```
test <- select(results, id, 1:9)
```

La función de selección es increíblemente útil para organizar datos rápidamente, como veremos a continuación.

13.19 Organizando los datos

Podríamos querer organizar nuestras filas (lo que puede ser bastante difícil de hacer en la base R). La arrange()función en dplyrfacilita esto y organiza los valores de una columna en orden ascendente de forma predeterminada. Aquí especificaremos la descendencia desc()porque queremos ver las publicaciones más recientes que mencionan nuestros términos de búsqueda en la parte superior.

```
results <- arrange(results, desc(publication_date))
  head(results$publication_date)
```

Cuando usemos View(results)veremos que los datos más recientes están en la parte superior. También veremos que algunos de los títulos en la parte superior son duplicados del mismo artículo porque incluyen todos los términos en nuestra búsqueda. Entonces, lo siguiente que querremos hacer es abordar los duplicados.

13.20 Tratando con duplicados

Cómo lidiar con los duplicados depende de lo que está tratando de lograr. Si está intentando desarrollar datos sobre tendencias, entonces los duplicados resultarán en un conteo excesivo a menos que tome medidas para contar solo registros distintos. Los duplicados de los mismos datos también distorsionarán la minería

de texto de las frecuencias de los términos. Entonces, desde esa perspectiva los duplicados son malos. Por otra parte. Si estamos interesados en el uso de términos a lo largo del tiempo dentro de un área emergente de la ciencia y la tecnología, entonces podríamos querer ver en detalle el uso de términos particulares. Por ejemplo, la genómica sintética es un término alternativo para la biología sintética favorecido por el grupo J. Craig Venter. Podríamos ver si este término es más ampliamente utilizado. ¿Los biólogos sintéticos también usan términos como biología de ingeniería, ¿La ingeniería del genoma o la nueva técnica de edición del genoma? En estos casos, los registros duplicados que usan términos son buenos porque los cambios en el idioma se pueden asignar a lo largo del tiempo. Esto sugiere la necesidad de una estrategia que use diferentes tablas de datos para responder a diferentes preguntas.

Como ya hemos visto, en R es muy fácil crear nuevos objetos (generalmente data.frames), realizar algún tipo de acción y escribir los datos en un archivo. Al pensar en los duplicados, es probable que primero deseamos averiguar con qué estamos tratando identificando registros únicos. Hay varias formas de hacer esto, aquí hay dos:

```
unique(results$id)  #displays unique DOIs (base R)
n_distinct(results$id)  #displays the count of distinct DOIs (dplyr)
```

Esto nos dice que hay 1,098 DOI únicos, lo que significa que había 307 duplicados en el momento de la escritura.

A continuación tenemos dos opciones principales.

1. Podemos difundir los resultados duplicados en la tabla.
2. Podemos identificar y borrar los duplicados.

13.20.1 Difundiendo datos usando spread()desdetidyr

En lugar de simplemente eliminar nuestros DOI duplicados, podríamos crear nuevas columnas para cada término de búsqueda y su DOI asociado. Esto será útil porque nos dirá qué términos están asociados con qué registros a lo largo del tiempo. Esto es fácil de hacer spread()proporcionando una keyy una valueen los argumentos. En este caso, queremos utilizarlos search_termscomo key(nombres de columna) para distribuirlos por la tabla y los DOI en la idcolumna como valuelas filas.

```
spread_results <- spread(results, search_terms, id)
```

Esto crea una columna para cada término de búsqueda con los DOI relevantes como valores. Tenga en cuenta que el valor predeterminado es eliminar la columna original (en este caso search_terms) al crear las nuevas columnas. Las cosas irán muy mal si intenta mantener la columna existente porque R intentará simultáneamente difundir los datos, reduciendo así el tamaño de la tabla y manteniendo la tabla en el mismo tamaño. Por lo tanto, dejaremos el valor predeterminado para eliminar la columna como está.

Ahora tenemos un data.frame con 1098 filas y los términos de búsqueda identificados en cada columna. Si examinamos brevemente spread_resultslos términos al final, podemos detectar un patrón potencialmente interesante donde algunos documentos solo usan términos como genoma sintético o genómica sintética, mientras que otros solo usan biología sintética o una combinación de términos.

Ahora hemos reducido nuestros datos a registros únicos a la vez que conservamos nuestros términos de búsqueda como puntos de referencia. La limitación de este enfoque es que al distribuir las DOI en 4 columnas ya no tenemos una columna ordenada de DOI.

13.20.2 Eliminando duplicados

Como alternativa, o complemento, para difundir podemos usar una prueba TRUE / FALSE lógica para filtrar nuestro conjunto de datos. Hay una serie de funciones que realizan pruebas lógicas en R (véase también which(), %in%, within()). En este caso la opción más adecuada es probablemente duplicated(). duplicated()marcará los registros duplicados como VERDADERO y los registros no duplicados como FALSO. Agregaremos una columna a nuestros datos usando la confianza $al crear la nueva columna.

```
results$duplicate <- duplicated(results$id)
```

Si usamos Ver (resultados), se agregará una nueva columna a los resultados. Los registros que no están duplicados se marcan como FALSO, mientras que los registros que están duplicados se marcan VERDADERO. Ahora queremos filtrar esa tabla a los resultados que no están duplicados (son FALSOS) de nuestra prueba lógica. Vamos a utilizar filter()desde dplyr(ver más arriba). Mientras que el select()trabajo exclusivo con columnas filter()funciona con filas y nos permite filtrar fácilmente los datos sobre los valores contenidos en una fila.

```
unique_results <- filter(results, duplicate == FALSE) %>%
    select(- search_terms)   #drop search_terms column
```

Aquí hemos pedido filter()que nos muestren solo aquellos valores en la columna duplicada que coincidan exactamente con FALSO. Ahora tenemos datos de 1097 resultados únicos con los DOI en una columna.

La creación de vectores lógicos TRUE / FALSE es muy útil para crear condiciones para filtrar datos. Sin embargo, en este caso, en la nota de proceso perderemos información de la search_termscolumna que quedará incompleta. Para evitar una posible confusión más adelante, eliminamos la search_termscolumna utilizando select(- search_terms)el código anterior. Si quisiéramos mantener los términos, usaríamos el método de propagación anterior.

Ahora tenemos tres data.frames, results, spread_results, y unique_results.

resultses nuestro núcleo o conjunto de referencia. Si planeamos hacer una cantidad significativa de trabajo con estos datos, guardaríamos una copia de results.csv y lo etiquetaríamos como rawcon notas en nuestro libro de códigos sobre sus orígenes y las acciones tomadas para generarlos. Puede ser una buena idea crear un .ziparchivo sin formato para que sea más difícil acceder por accidente.

En el futuro usaríamos el spread_resultsy unique_resultspara trabajo adicional.

Como hicimos anteriormente, use write.csv (x, "x.csv", row.names = FALSE) o el más simple y más rápido write_csv(). R puede escribir varios archivos en un abrir y cerrar de ojos. Esto escribirá los tres archivos en la carpeta del proyecto rplos (use getwd()y setwd()si quiere hacer algo diferente).

```
write_csv(results, "results.csv")
  write_csv(spread_results, "spread_results.csv")
  write_csv(unique_results, "unique_results.csv")
```

Ok, ahora tenemos un conjunto de datos que contiene los registros de un rango de términos y hemos recorrido un largo camino. Mucho de esto ha sido sobre qué hacer con los datos de PLOS una vez que hemos accedido a ellos en términos de convertirlos en tablas con las que podamos trabajar. En la siguiente sección veremos cómo restringir las búsquedas por sección.

13.21 Restricción de búsquedas por sección

El valor predeterminado para buscar rploses buscar todo. Esto puede producir muchos resultados que pasan y ser abrumador. Hay bastantes opciones para restringir las búsquedas en rplos.

13.22 por el autor

Al crear el conjunto de datos de resultados anterior, incluimos el authorcampo. Sin embargo, hay algunas complejidades para buscar con nombres de autores y trabajar con datos de autores que es importante comprender. Comenzaremos buscando nombres de autores y luego veremos cómo procesar los datos.

Para restringir una búsqueda por nombre de autor podemos usar el nombre completo o el apellido:

```
plosauthor(q = "Paul Oldham", fl = c("author", "id"),
 fq = "doc_type:full",       limit = 20)
```

En este ejemplo, hemos especificado doc_type:fulldevolver solo los resultados de los artículos completos. Si no usa esto, la búsqueda devolverá una gran cantidad de resultados repetidos en función de las secciones del artículo. Entonces, en este caso, Paul Oldham, el autor de este artículo rplos, ha publicado dos artículos en PLOS ONE. Si doc_type:fullno se especifica, se devuelven más de 20 resultados que muestran diferentes secciones de los dos artículos. Esto creará un problema de duplicación más adelante, por lo que se debe usar un enfoque predeterminado razonable doc_type:full.

Como observación general, se debe tener mucho cuidado al trabajar con nombres de autores debido a problemas con el agrupamiento de nombres y la división de nombres como se describe en este artículo de PLOS ONE . Si se encuentra un gran número de resultados en un solo nombre de autor, considere utilizar criterios de coincidencia de otros campos de datos disponibles para asegurarse de que no se agrupen personas separadas por nombre. Sobre todo, no asuma que simplemente porque un nombre es el mismo, o muy similar al nombre de destino, el nombre designa a la misma persona.

El siguiente problema que debemos abordar es qué hacer con los datos del autor cuando los hayamos recuperado. La razón de esto es que el campo del autor en los resultados es generalmente un campo concatenado que contiene los nombres de los autores de un artículo en particular. Comenzaremos con el oldhamconjunto de resultados.

En este caso, realizaremos la llamada plosauthor()y luego utilizaremos ldply()desde plyrpara devolver un marco de datos que contenga metay data. Luego usaremos filldesde tidyrpara tomar numFoundy completar esa columna. Eliminaremos la columna de inicio usando select()y finalmente filter()limitaremos la tabla a los datos.

```
oldham <- plosauthor(q = "Paul Oldham", fl = c("author", "id"),
fq = "doc_type:full", limit = 20) %>%    ldply("[", 1:2) %>%
    fill(numFound, start) %>%    select(- start) %>%
    filter(.id == "data")
```

Ahora tenemos dos registros con los datos de autor e ID (DOI). Lo siguiente que queremos hacer es separar los nombres de los autores. Podemos hacer esto usando separate(). Tenga en cuenta que separate()deberá saber la cantidad de nombres involucrados de antemano. En el caso de datos de Oldham hay tres autores de cada artículo. Nos ocuparemos de cómo calcular el número de nombres de autores en breve.

```
oldham <- separate(oldham, author, 1:3, sep = ";", remove = FALSE)
```

Ahora tenemos algunas otras opciones. Simplemente podríamos mantener solo el nombre del primer autor. Para hacer eso, en este caso particular, podríamos usar select()y la posición numérica de las columnas que queremos eliminar.

```
first_author <- select(oldham, -7, -8)
```

Como alternativa, podríamos colocar cada nombre de autor en su propia fila para que podamos enfocarnos en un autor específico más adelante. Para eso podemos usar gather()desde tidyry los números de posición de columna (no sus nombres en este caso) de las columnas que queremos reunir.

```
authors <- gather(oldham, number, authors, 5:7)
```

Lo anterior gather()requiere un campo de clave y valor. En este caso, hemos utilizado el número como nuestra clave y los autores como nuestro valor. Luego, hemos especificado que queremos reunir las columnas 6 a 8 en los nuevos autores de columna.

Eso fue fácil porque estamos tratando con un pequeño número de resultados con un número uniforme de autores. Sin embargo, nuestros resultsdatos son más complicados que esto porque tenemos varios nombres de autores para cada artículo y el número de autores para los artículos podría variar considerablemente.

Necesitaremos organizar los datos y realizar algunos cálculos simples para que esto funcione. Esto llevará seis pasos. El código de trabajo completo está abajo.

1. Calculamos el número de columnas en nuestro conjunto de datos. Hacemos esto porque el número puede variar dependiendo de los campos que recuperamos rplos. Vamos a utilizar ncols()para hacer el cálculo.

2. Usamos una función corta de stringrpara calcular el número de autores según el separador de nombre del autor ";" (+1 para capturar los nombres finales en la secuencia). Esto nos da el número máximo de autores en el conjunto de datos en el que necesitamos dividir los datos (en este caso, 83 como el valor de n). Copia y pega la siguiente función en la consola para acceder a ella.

```
author_count <- function(data, col = "", sep = "[^[:alnum:]]+") {
     library(stringr)
      authcount <- str_count(data[[col]], pattern = sep)
      n <- as.integer(max(authcount) + 1)      print(n)  }
```

3. Usamos select()desde dplyrpara mover nuestra columna de destino a la primera columna. Esto simplemente hace que sea más fácil especificar las posiciones de las columnas separate()y gather()más adelante.
4. Usamos el valor de npara separar los nombres de los autores en varias columnas.
5. Luego los reunimos de nuevo usando el valor de n.
6. La división en un separador como el que ;normalmente genera espacios en blanco iniciales y finales invisibles. Esto evitará que los nombres de los autores se clasifiquen correctamente (por ejemplo, en Excel o Tableau). La str_trim()función de stringrproporciona una manera fácil de eliminar el espacio en blanco (especifique el lado como derecho, izquierdo o ambos).

Copie y pegue el siguiente código y luego presione Enter.

```
#---calculations---  colno <- ncol(unique_results)
  #calculate number of columns
  n <- author_count(unique_results, "author", ";")
  # See function above.
 Calculate n as an integer to meet requirement for separate()
  #---select, separate and gather---
 full_authors <- select(unique_results, author, 1:colno)
 #bring author to the front
  full_authors <- separate(full_authors, author, 1:n, sep = ";",
 remove = TRUE,
     convert = FALSE, extra = "merge", fill = "right")  #separate
 full_authors <- gather(full_authors, value, authors, 1:n, na.rm = TRUE)
  #gather  #---trim authors----
full_authors$authors <- str_trim(full_authors$authors, side = "both")
```

```
  #trim leading and trailing whitespace
```

Podemos simplificar esto con tuberías para reunir las acciones en el nuevo objeto full_authors.

```
#---calculations---  colno <- ncol(unique_results)
  n <- author_count(unique_results, "author", ";")
  #---select, separate, gather---
  full_authors <- select(unique_results, author, 1:colno)
 %>% separate(author,
       1:n, sep = ";", remove = TRUE, convert = FALSE, extra = "merge",
 fill = "right") %>%       gather(value, authors, 1:n, na.rm = TRUE)  #--
-trim authors----
 full_authors$authors <- str_trim(full_authors$authors, side = "both")
```

Al ejecutar este código, eliminaremos la columna del autor original (columna 1) especificando remove = TRUEen separate(). gather()Colocará la nueva authorscolumna al final. Por lo tanto, asegúrese de desplazarse a la columna final cuando vea los resultados. También podríamos caer columnas no deseadas.

Ahora tenemos una lista completa de nombres de autores individuales que podrían usarse para buscar autores individuales, para limpiar nombres de autores para uso estadístico y para mapeo de red de autores. Como un breve ejemplo, si quisiéramos buscar las contribuciones de Jean Peccoud que dirige el blog PLOS SynBio , podríamos usar lo siguiente en función de esta útil respuesta de desbordamiento de pila . Ver? Grepl para más información.

```
Peccoud <- filter(full_authors, grepl("Peccoud", authors))
```

No profundizaremos en estos temas, pero generar este tipo de lista de autores es un paso importante para permitir un análisis y una visualización más amplios. Si bien el código utilizado para llegar a esta lista de autores puede parecer bastante complicado, una vez que se entienden los conceptos básicos, se puede usar una y otra vez.

Vamos a escribir esos datos en un archivo .csv para explorar más adelante.

```
write_csv(full_authors, "full_authors.csv")
```

13.23 Búsqueda de título usandoplostitle()

Para una búsqueda de título podemos usar plostitle(). Como se indica anteriormente, es posible que desee contar primero el número de registros utilizando:

```
t <- plostitle(q = "synthetic biology", limit = 0)$meta$numFound
```

Luego ejecutamos la búsqueda para devolver el número de resultados que nos gustaría. Aquí lo hemos puesto al valor de t arriba (11). Hemos limitado los resultados al campo de datos por subconjunto con $ datos.

```
title <- plostitle(q = "synthetic biology", fl = "title", limit = t)
$data
```

13.24 búsqueda abstracta usandoplosabstract()

Para limitar las búsquedas a los resúmenes que podemos utilizar plosabstract(). Comenzaremos con un conteo rápido de registros.

```
a <- plosabstract(q = "synthetic biology", limit = 0)$meta$numFound
```

Para recuperar los resultados podríamos usar el valor de a. Como alternativa, podríamos establecerlo arbitrariamente alto y se devolverán los resultados correctos. Por supuesto, si no sabemos cuál es el número total de resultados, entonces no estaremos seguros de si hemos capturado el universo. Pero, un número arbitrario puede ser útil para la exploración.

```
abstract <- plosabstract(q = "synthetic biology", fl = "id, title,
 abstract",       limit = 200)  abstract$data
```

Como antes, podemos crear fácilmente un nuevo objeto que contenga el data.frame. En este caso también se incluyen los metadatos y luego usar fill()de tidyr()llenar por el numFoundcampo y el comienzo con 0. Obsérvese que metaaparecerá en la parte superior de la lista y creará una fila en gran parte en blanco. Para evitar esto, mientras mantenemos el número de registros para referencia, usaremos el filtro de tidyr(). Este código corto lo hará.

```
abstract_df <- ldply(abstract, "[", 1:2) %>% fill(numFound, start)
 %>% filter(.id ==       "data")
```

13.25 Área temática utilizandoplossubject()

Para buscar por área temática utilizar plossubject. El retorno predeterminado es 10 resultados de los resultados totales. Entonces, intente comenzar con una búsqueda como esta para tener una idea de cuántos resultados hay. En este caso, la consulta se ha limitado a PLOS ONE y artículos de texto completo.

```
sa <- plossubject(q = "\"synthetic+biology\"",
 fq = list("cross_published_journal_key:PLoSONE",
      "doc_type:full"))$meta$numFound
```

Al momento de escribir esto se devuelven 739 resultados. Simplemente retiraremos 10 resultados. Para retirar todos los resultados, reemplace 10 por saarriba o escriba el número en limit =.

```
plossubject(q = "\"synthetic+biology\"",
 fl = "id", fq = list("cross_published_journal_key:PLoSONE",
       "doc_type:full"), limit = 10)
```

Como se señala en la documentación, los resultados que devolvemos de la API y los resultados en el sitio web no son necesariamente los mismos porque la configuración utilizada por PLOS en el sitio web no está clara.

En este caso, devolvemos 740 resultados, mientras que, en el momento de redactar este informe, PLOS ONE enumera 417 artículos en el área temática de Biología sintética . Esto merecerá una aclaración de los criterios para los recuentos utilizados en el sitio web de PLOS y las devoluciones de API.

13.26 Resaltando términos y fragmentos de texto con highplos ()

highplos()es una gran función para la investigación en PLOS, particularmente cuando se combina con abrir resultados en un navegador usando highbrow().

El resaltado hará retroceder una parte del texto con el término de búsqueda resaltado con la etiqueta de énfasis que encierra las palabras individuales en una frase de búsqueda. Es posible que se pueda resaltar una frase completa (vea hl.usePhraseHighlighter) pero esto requiere una mayor exploración.

En este ejemplo, simplemente usaremos el término biología sintética y luego resaltaremos los términos en el resumen hl.fl =y lo limitaremos a 10 filas de

resultados. También agregaremos la función highbrow()(para resaltar navegar) al final. Esto abrirá los resultados en nuestro navegador. En los ejemplos usamos una tubería (%>%) que significa this %then% that. Esto significa que no tenemos que ingresar el fragmento de nombre en la función highbrow y simplifica el código.

Al revisar los resultados en un navegador, tenga en cuenta que podemos hacer clic en el DOI para ver el artículo completo. Esta es una herramienta realmente útil para evaluar qué artículos queremos analizar más detenidamente.

```
highplos(q = '"synthetic biology"', hl.fl = 'abstract',
fq = "doc_type:full", rows = 10) %>%
  highbrow() #launches the browser
```

Tenga en cuenta que en algunos casos, a pesar de que estamos restringiendo doc-type:full, recuperamos entradas sin datos. En un caso, esto se debe a que estamos resaltando los términos en el resumen cuando el término aparece en el texto completo. En un segundo caso, hemos seleccionado una corrección en la que uno de los autores se encuentra en un centro de biología sintética, pero ni el resumen ni el texto mencionan la biología sintética. Por lo tanto, tenga en cuenta que es posible que se requiera una exploración adicional para comprender por qué se devuelven resultados particulares. Estos problemas son menores y esta es una gran herramienta.

Hay dos opciones adicionales (argumentos) highplos()que podemos usar. El primero de estos es fragmentos utilizando hl.snippets =y el segundo es hl.fragsize =. Ambos se pueden utilizar en conjunto con highbrow().

13.26.1 Fragmentos usando hl.snippets

```
snippet <- highplos(q = '"synthetic biology"', hl.fl = list("title",
 "abstract"), hl.snippets = 10, rows = 100) %>%    highbrow()
```

El argumento de los fragmentos de código es útil (el valor predeterminado para un fragmento de código es 1, pero sube a tantos como desee). Se vuelve muy interesante cuando añadimos hl.mergeContiguous = 'true'. Esto mostrará las entradas capturadas en el orden de los artículos para proporcionar un sentido de sus usos por el autor (es).

```
highplos(q='"synthetic biology"', hl.fl = "abstract", hl.snippets = 10,
 hl.mergeContiguous = 'true', rows = 10) %>%    highbrow()
```

13.26.2 tamaño de fragmento usando hl.fragsize

Un mayor control sobre lo que estamos viendo se proporciona mediante la hl.fragsizeopción. Esto nos permite especificar el número de caracteres (incluidos los espacios) que queremos ver en relación con nuestros términos de destino.

En el primer ejemplo, resaltaremos la frase biología sintética en los títulos y resúmenes y estableceremos el tamaño del fragmento (usando hl.fragsize) a un máximo de 500. Esto devolverá los primeros 500 caracteres, incluyendo espacios en lugar de palabras. Estableceremos el número de filas en un valor algo arbitrario de 200. Esto puede empujarse fácilmente mucho más alto, pero espere un momento si lo mueve a 1000 filas.

```
highplos(q = '"synthetic biology"', hl.fl = list("title", "abstract"),
 hl.fragsize = 500, rows = 200) %>%    highbrow()
```

También podemos hacer lo contrario de una búsqueda más grande al reducir el tamaño del fragmento para decir hasta 100 caracteres. En este momento no está claro si es posible controlar si los caracteres se seleccionan a la derecha o a la izquierda de nuestros términos de destino. Tenga en cuenta que los resultados mostrarán hasta 100 caracteres cuando estén disponibles (los resultados cortos serán para oraciones, como títulos que tienen menos de 100 caracteres)

```
highplos(q = '"synthetic biology"', hl.fl = list("title", "abstract"),
 hl.fragsize = 100, rows = 200) %>%    highbrow()
```

Lo bueno de esto es que podemos controlar fácilmente la cantidad de texto que estamos viendo y luego seleccionar artículos de interés para leer directamente desde el navegador. También podemos empezar a pensar en formas de usar esta información para la minería de textos para identificar términos utilizados en conjunto con biología sintética o tipos de biología sintética.

13.27 Obtenga el texto completo de uno o más artículos

Terminaremos este artículo demostrando brevemente cómo recuperar y guardar el texto completo de uno o más artículos. rplosutiliza una combinación del paquete XMLy tm(para la minería de texto).

La recuperación del texto completo debe usarse inicialmente con moderación, ya que podría recuperar una gran cantidad de datos en formato XML que luego podría tener problemas para procesar. Por lo tanto, probablemente es mejor comenzar poco a poco.

Usando los datos de unique_results que creamos anteriormente, tenemos una lista de DOI en el campo id. Podemos crear un vector de estos usando lo siguiente:

```
doi <- unique_results$id
```

Eso ha creado un vector de 1097 dois. Para limitar esos resultados, vamos a crear una versión más corta donde seleccionamos cinco filas.

```
short_doi <- doi[1:5]
```

Ahora podemos utilizarlo plos_fulltext()para recuperar el texto completo.

```
ft <- plos_fulltext(short_doi)
```

Cuando retiramos los dos artículos, se crea un objeto de clase plosft. Para ver el texto completo de uno de los artículos individuales, usamos el confiable $y luego seleccionamos un doi.

```
ft$`10.1371/journal.pone.0140969`
```

Esto muestra muchas de las etiquetas XML dentro del texto. Ahora nos gustaría extraer el texto sin las etiquetas XML. La rplosdocumentación para plos_fulltext()nos ayuda a hacer esto usando el siguiente código. La primera parte del código utiliza el paquete XML para analizar los resultados eliminando las etiquetas xml en el proceso.

```
library(tm)
  library(XML)
  ft_parsed <- lapply(ft, function(x) {
      xpathApply(xmlParse(x), "//body", xmlValue)  })
```

Si escribimos ft_parsed, veremos que el texto (el cuerpo sin título y el resumen) pasa volando sin todas las etiquetas.

```
ft_parsed
```

El objeto devuelto por esto es una lista (uso class(ft_parsed)). A continuación, podemos transformar esto en un corpus (un texto o colección de textos) que podemos guardar en el disco usando el siguiente código del rplos plos_fulltext()ejemplo.

```
tmcorpus <- Corpus(VectorSource(ft_parsed))
```

Si escribimos tmcorpus $ en la consola, veremos 1 a 5 emergente, pero esto devolverá NULL si se selecciona. Los datos están ahí, pero necesitamos usarlos str(tmcorpuspara ver la estructura del corpus. Si queremos ver un texto dentro del corpus podemos usarwriteLines()

```
writeLines(as.character(tmcorpus[[2]]))
```

También podemos ver los cinco textos de nuestro corpus (prepárese para una gran cantidad de desplazamiento) utilizando lapply para leer los dos textos como caracteres.

```
lapply(tmcorpus[1:5], as.character)
```

Para obtener más información, consulte la Introducción de Ingo Feinerer (2015) al paquete tm (también disponible en la documentación de tm) del cual se extrae lo anterior.

13.28 Escribiendo un corpus al disco.

Para escribir un corpus primero debemos crear una carpeta donde se alojarán los archivos (de lo contrario, simplemente se escribirán en la carpeta de su proyecto con todo lo demás).

La forma más fácil de crear una nueva carpeta es dirigirse a la pestaña Archivos en RStudio (normalmente en el panel inferior derecho) y elegir New Folder. Lo llamaremos tm.

Ahora use getwd()y copie la ruta del archivo en la siguiente función, de los ejemplos de writeCorpus, agregando /tmal final. Se verá algo como esto pero reemplazará el camino por el tuyo, sin olvidar el /tm. Luego presione Enter.

```
writeCorpus(tmcorpus,
 path = "/Users/paul/Desktop/open_source_master/rplos/tm")
```

Cuando busque en la carpeta tm dentro de rplos (use la pestaña Archivos en RStudio) ahora verá cinco textos con los nombres del 1 al 5. Para más detalles, como nombrar archivos y especificar tipos de archivos, consulte ?writeCorpusy la documentación del paquete tm.

13.29 Round Up

En este capítulo nos hemos centrado en utilizar el rplospaquete para acceder a artículos científicos de la Biblioteca Pública de Ciencias (PLOS). Como hemos

visto, con piezas cortas de código es fácil buscar y recuperar datos de PLOS en una amplia gama de temas, ya sea pizza o biología sintética.

Una de las características más poderosas de R es que es bastante fácil acceder a datos en línea gratuitos utilizando APIs. rploses un muy buen punto de partida para aprender a recuperar datos usando una API porque está bien escrito y los datos que regresan son notablemente limpios.

Quizás el mayor desafío al que se enfrentan los nuevos usuarios de R es qué hacer con los datos una vez que los haya recuperado. Esto puede resultar en muchas horas de frustración al mirar una lista u objeto con los datos que necesita sin las herramientas para acceder y transformarlos en el formato que necesita. En este artículo nos hemos centrado en el uso de plyr, dplyr, tidyry el stringrconjunto de paquetes para convertir rplosdatos en algo que puede utilizar. Estos paquetes son, con razón, muy populares para el trabajo diario en R y familiarizarse con ellos generará recompensas en el aprendizaje de R para el trabajo práctico. Al final del artículo, utilizamos el tmpaquete (extracción de texto) para guardar el texto completo de los artículos. Esto es solo una parte muy pequeña de este paquete y rplosproporciona algunos ejemplos útiles para comenzar la minería de texto usandotm(ver los plos_fulltext()ejemplos). R ahora tiene una amplia gama de paquetes de minería de texto y lo abordaremos en un próximo artículo.

Mientras tanto, si desea obtener más información sobre R, pruebe los siguientes recursos. Si desea aprender R dentro de R, pruebe el muy útil paquete Swirl (detalles a continuación).

13.30 recursos

1. rOpenSci
2. El libro de cocina R de Winston Chang
3. Aprendizaje en línea RStudio
4. r-bloggers.com
5. Datacamp
6. Swirl (desarrollado por el equipo gratuito del curso de Programación Coursea R en la Universidad John Hopkins. Si desea comenzar con Swirl, ejecute el fragmento de código a continuación para instalar el paquete y cargar la biblioteca.

```
install.packages("swirl")  library(swirl)
```