

Introducción a *Voyant Tools*:

Análisis textual de *General Estoria I*

Introducción:

Esta guía proporciona una introducción a *Voyant Tools*, una aplicación de código abierto para realizar minería de texto y datos. Desarrollada por Stéfán Sinclair de McGill University y Geoffrey Rockwell de University of Alberta, *Voyant Tools* fue creada para apoyar la lectura e interpretación académica de textos.

Creada pensando en los estudiosos de las humanidades digitales, *Voyant Tools* proporciona listas de frecuencia de palabras, gráficos de distribución de frecuencia y análisis KWIC (Key Word in Context). Para obtener más información sobre *Voyant Tools*, visite el [repositorio de Voyant Tools](#) en GitHub o la [Guía de ayuda](#) de *Voyant Tools*.

Antes de empezar:

Se puede acceder a *Voyant Tools* en línea en <https://voyant-tools.org>. También se puede instalar el servidor Voyant como una versión independiente en nuestro propio ordenador. Esto tiene varias ventajas potenciales, incluyendo un mejor rendimiento, fiabilidad, seguridad y privacidad de los datos analizados.

Antes de instalar el servidor Voyant en el ordenador, es preciso tener Java instalado y configurado en nuestro ordenador. Después hay que acceder la página con las [últimas versiones](#) de *Voyant Tools* y hacer clic en el archivo [VoyantServer2.6.13.zip](#) para descargar los archivos necesarios para configurar el servidor. Se trata de un archivo zip de gran tamaño, unos 200 MB, que incluye grandes modelos de datos para el procesamiento del lenguaje. Será necesario descomprimir el archivo zip antes de poder instalarlo.

En el [Repositorio de Github del Servidor Voyant](#) podemos encontrar información e instrucciones detalladas sobre la instalación del servidor Voyant. Para su correcto funcionamiento Voyant Server necesita que Java 11 esté instalado en nuestro ordenador (PC o Apple). Una vez descomprimido el archivo [VoyantServer2.6.13.zip](#), solo es necesario hacer clic en `VoyantServer.jar` para ejecutar la aplicación.¹

Formatos de archivo aceptados:

[Voyant Tools](#) es un entorno de lectura y análisis de texto basado en la web que permite cargar textos en diferentes formatos—TXT, HTML, XML, PDF, RTF y MS Word—para su posterior análisis. Es posible crear nuestra propia colección de textos, o utilizar uno de los corpus de muestra disponibles en *Voyant Tools*.

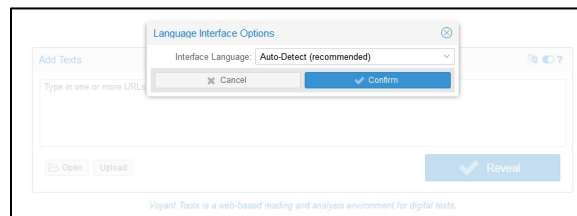
¹ En algunos casos, si el programa no se ejecuta en los ordenadores Apple, es necesario abrir la terminal, cambiar al directorio de VoyantServer, `cd VoyantServer`, y escribir `/Library/Java/JavaVirtualMachines/jdk-11.jdk/Contents/Home/bin/java -jar VoyantServer.jar`

Cómo cambiar el lenguaje de la interfaz de *Voyant Tools*

Para cambiar el lenguaje de la interfaz, hay que hacer clic en el botón señalado con un círculo rojo



y luego seleccionar el lenguaje en el menú desplegable:



Es también posible cambiar el lenguaje de la interfaz añadiendo `&lang=es` al final del URL:

`http://127.0.0.1:8888/?corpus=c6f199d18226c6acd13b8fee3a494eac&lang=es`

Cómo subir textos a *Voyant Tools*

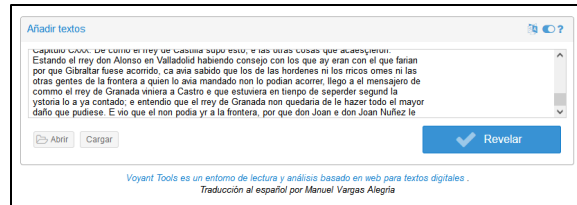
Hay cuatro formas de subir textos a *Voyant Tools*:

1. Abriendo un corpus existente con *Voyant Tools*: Haciendo clic en el botón **Abrir** debajo del cuadro de texto y seleccionando uno de los corpus de muestra proporcionados por *Voyant Tools*.

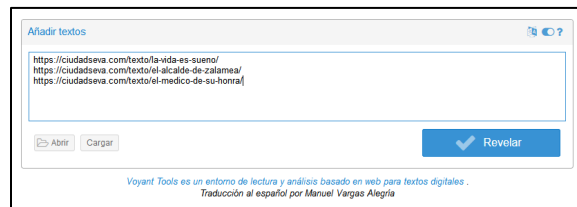


Voyant Tools incluye tres corpus de muestra: “Shakespeare’s Plays”, “Austen’s Novels” y “Mary Shelley’s *Frankenstein*”. Estos corpus fueron creados a partir de los textos del [Proyecto Gutenberg](#). El corpus, “Shakespeare’s Plays” incluye todas las 37 obras de William Shakespeare. El corpus “Austen’s Novels” incluye los textos de ocho de las obras de Jane Austen, y “Mary Shelley’s *Frankenstein*” el texto completo de la obra.

2. Escribiendo o pegando texto en el cuadro de texto principal (esto crea un corpus con un único documento)



3. Escribiendo o pegando una o más URL en el cuadro de texto principal (una URL por línea)



4. Haciendo clic en el botón **Cargar** debajo del cuadro de texto para cargar archivos desde nuestro ordenador. Hay que seleccionar todos los archivos a la vez (pulsar la tecla Shift en el teclado del ordenador y hacer clic en cada elemento) o comprimir los archivos en una carpeta comprimida y cargar el archivo zip en *Voyant Tools*.

Tutorial: Análisis textual de *General Estoria I*

Ahora que hemos repasado cómo acceder a las herramientas de *Voyant Tools* y subir textos, podemos examinar nuestro propio corpus. En este tutorial, usaremos *Voyant Tools* para observar las frecuencias de palabras, las colocaciones, las tendencias y las palabras clave en contexto en un corpus formado por los 29 libros que componen la *General Estoria* I. Este corpus fue obtenido de la transcripción paleográfica del Hispanic Seminary of Medieval Studies.

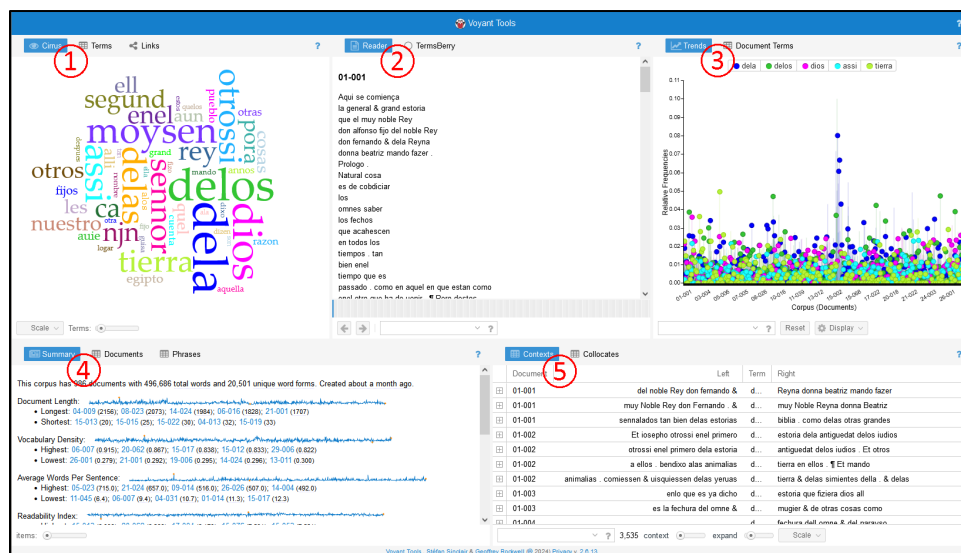
Paso 1: Cómo cargar un corpus

Primero, hay que descargar el fichero GE1-libros.zip en el ordenador. Este archivo zip contiene nuestro corpus de muestra, compuesto por los 29 libros de la *General Estoria I* en archivos de texto plano (txt).

Una vez conectados a *Voyant Tools*, hacer clic en subir y seleccionar el archivo GE1-libros.zip que acabamos de guardar en el ordenador y hacer clic en **Revelar**. El contenido del archivo zip debería cargarse automáticamente dentro de la interfaz de *Voyant Tools*.

Paso 2: La interfaz de trabajo







La interfaz de *Voyant Tools* muestra cinco paneles con diferentes herramientas de análisis de texto.



Las siguientes herramientas de análisis de texto están disponibles por defecto:

1. **Cirrus:** visualiza en forma de nube de palabras aquellas de mayor frecuencia de un corpus o documento. La nube de palabras sitúa las palabras de tal manera que los términos que aparecen con más frecuencia aparecen en el centro y tienen el tamaño más grande. A medida que el algoritmo recorre la lista y continúa dibujando palabras lo más cerca posible del centro de la visualización, también incluirá palabras pequeñas dentro de los espacios dejados por palabras más grandes que no encajan perfectamente. Es importante entender que el color de las palabras y su posición absoluta no son significativos (si cambiamos el tamaño de la ventana o volvemos a cargar la página, las palabras pueden aparecer en una ubicación diferente).

2. **Lector:** muestra el texto para su lectura. Es posible desplazarse hacia abajo dentro del lector de texto para obtener más contenido. También al pasar el cursor sobre una palabra se muestra su frecuencia en el documento. Además, se puede hacer clic en una palabra o buscarla en el cuadro de búsqueda para ver con qué frecuencia aparece en el corpus.
3. **Tendencias:** muestra el gráfico de distribución que representa las frecuencias de los términos en los textos de su corpus. Cada serie del gráfico está coloreada según la palabra que representa. En la parte superior del gráfico, una leyenda muestra qué palabras están asociadas con ciertos colores. Se puede hacer clic en las palabras de la leyenda para cambiar su visibilidad. Al mantener el puntero sobre cualquier punto del gráfico, aparece un cuadro de llamada con información sobre el punto, incluida la palabra y su frecuencia.
4. **Sumario:** ofrece un resumen de las características del corpus, incluyendo el número de documentos y el número total de palabras (tokens/forms) y palabras únicas (múltiples apariciones de palabras) en el corpus. La siguiente parte del resumen muestra la longitud del documento del corpus. Muestra los documentos más largos y más cortos por el número de palabras del corpus. Entre paréntesis, después de cada título del corpus, aparece el número de palabras. La siguiente sección muestra los documentos con las densidades de vocabulario más altas (la relación entre el número de palabras del documento y el número de palabras únicas del documento) y los documentos con las densidades de vocabulario más bajas. A continuación de esta sección se muestra una aproximación del número medio de palabras por frase, tanto los valores más altos como los más bajos. A continuación, aparecen las cinco palabras más frecuentes en el corpus, las mismas que aparecen en **Tendencias**.
5. **Contextos:** muestra una concordancia KWIC (palabra clave en contexto). La tabla tiene las siguientes columnas:
 - **Documento:** muestra en qué documento se encuentra la palabra clave y los contextos
 - **Izquierda:** palabras contextuales a la izquierda de la palabra clave (la clasificación por esta columna trata las palabras en orden inverso, de derecha a izquierda de la palabra clave)
 - **Términos:** la palabra clave que coincide con la consulta de términos predeterminada o proporcionada por el usuario
 - **Derecha:** palabras contextuales a la derecha de la palabra clave

Para seleccionar una herramienta alternativa, es preciso pasar el cursor sobre la barra gris en la parte superior de la ventana de **Tendencias** o **Cirrus** junto al símbolo  hasta que aparezca un menú de iconos    . Al seleccionar el botón de ventana  aparece un listado de todas las otras herramientas que pueden realizar diferentes visualizaciones y análisis de texto. En **Paso 7: Otras herramientas** se explica el funcionamiento de varias de estas herramientas.

Paso 3: Nubes de palabras y palabras vacías

Nube de palabras

La herramienta **Cirrus** muestra una nube de palabras en la que aparecen las palabras más frecuentes en

[illegible]

Herramienta de resumen

SummaryDocumentsPhrases?

Most frequent words in the corpus:

- [ball](#) (3535), [kale](#) (3119), [sax](#) (2814), [box](#) (2342), [lame](#) (2254)

Distinctive words (compared to the rest of the corpus):

- 01-001: [brat](#) (2), [kale](#) (4), [herman](#) (2), [lecter](#) (9), [començau](#) (2).
- 01-002: [era](#) (5), [lar](#) (3), [crepessan](#) (2), [archeviesse](#) (2), [ague](#) (5).
- 01-003: [beteno](#) (3), [kreasau](#) (2), [maternal](#) (1), [insolera](#) (1), [aportuna](#) (1).
- 01-004: [parayse](#) (8), [començau](#) (5), [ball](#) (5), [dilecto](#) (3), [fala](#) (3).
- 01-005: [parayse](#) (11), [letra](#) (17), [sax](#) (10), [sede](#) (4), [tumpai](#) (3).
- 01-006: [sua](#) (8), [adare](#) (6), [layn](#) (3), [parayse](#) (3), [salterri](#) (2).
- 01-007: [abel](#) (7), [leibora](#) (3), [commo](#) (4), [ague](#) (2), [haha](#) (2).
- 01-008: [abel](#) (6), [layn](#) (3), [layn](#) (2), [abell](#) (2), [marchisti](#) (1).
- 01-009: [abel](#) (10), [layn](#) (8), [volensia](#) (4), [sanea](#) (3), [sacrefica](#) (3).
- 01-010: [layn](#) (13), [abel](#) (5), [herman](#) (6), [sax](#) (6), [guarall](#) (2).
- Next 10 of 976 remaining

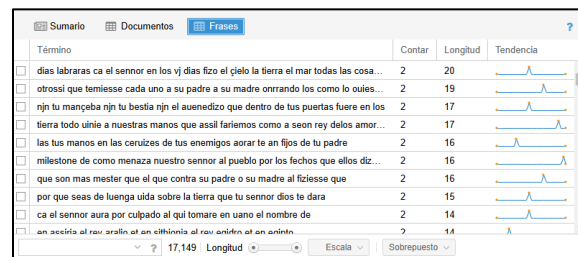
	Título	Palabras	Tipos	Proporción	Palabras/oración
1	01	12.375	2.074	17%	44.8
2	02	17.194	2.625	15%	44.3
3	03	18.494	2.700	15%	58.7
4	04	17.524	2.810	16%	85.5
5	05	20.247	3.226	16%	94.6
6	06	21.134	3.281	16%	56.8
7	07	23.566	3.409	14%	66.4
8	08	24.125	3.375	14%	88.4
9	09	18.238	2.695	15%	97.5
10	10	13.710	1.976	16%	171.6

- **Título:** el título del documento (o el nombre del archivo si no se ha encontrado un título mejor).
- **Palabras:** el número de palabras individuales (tokens) encontradas en el documento (por

ejemplo, se cuenta cada aparición de "assi")

- **Tipos:** número de formas de palabras encontradas en el documento (por ejemplo, todas las apariciones de "assi" se cuentan como una forma de palabra).
- **Proporción:** relación entre los tipos y las palabras individuales (tokens), expresada en porcentaje. Un número más alto suele significar una mayor diversidad de vocabulario.
- **Palabras/frase:** una aproximación del número medio de palabras por frase (recuento de palabras/frases); la forma en que se calculan las frases debe considerarse muy aproximada, especialmente debido a las complicaciones con las abreviaturas y otros usos de la puntuación (el análisis sintáctico de las frases lo realiza la clase BreakIterator de Java, y también depende de la detección precisa del idioma).

La herramienta **Frases**, por su parte, muestra secuencias repetidas de palabras organizadas por frecuencia de repetición o número de palabras en cada frase repetida. Por defecto, las frases se muestran en orden descendente de longitud de frase.



Término	Contar	Longitud	Tendencia
<input type="checkbox"/> días labraras ca el sensor en los vj días fizo el cielo la tierra el mar todas las cosa...	2	20	
<input type="checkbox"/> otrossi que temiesse cada uno a su padre a su madre onrrando los como lo oules...	2	19	
<input type="checkbox"/> njn tu mançeba njn tu bestia njn el auenedizo que dentro de tus puertas fuere en los	2	17	
<input type="checkbox"/> tierra todo unie a nuestras manos que assil fariemos como a seon rey delos amor...	2	17	
<input type="checkbox"/> las tus manos en las ceruizes de tus enemigos aorar te an fijos de tu padre	2	16	
<input type="checkbox"/> milestone de como menaza nuestro sensor al pueblo por los fechos que ellos diz...	2	16	
<input type="checkbox"/> que son mas mester que el que contra su padre o su madre al fizesse que	2	16	
<input type="checkbox"/> por que seas de luenga uida sobre la tierra que tu sensor dios te dara	2	15	
<input type="checkbox"/> ca el sensor aura por culpado al qui tomare en uano el nombre de	2	14	
<input type="checkbox"/> en acciito al ray aralin al an olliñito al ray aralin al an arinle...	2	14	

17,149 Longitud Escala Sobrepuesto

La herramienta **Frases** muestra cuatro columnas de forma predeterminada:


- **Término:** la frase que se repite.
- **Contar:** el número de veces que la frase aparece en el documento.
- **Longitud:** el número de palabras de la frase.
- **Tendencia:** el gráfico que muestra la distribución de frecuencias relativas del término en el documento.

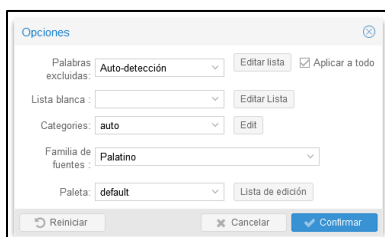
La barra inferior de la herramienta **Frases** tiene una barra de búsqueda, y varias opciones adicionales.

- **Barra de búsqueda:** permite buscar términos concretos y qué frases contienen dicho término, o buscar determinadas frases en conjunto.
- **Escala:** el control deslizante señala los límites superior e inferior de la longitud de la frase.
- **Sobrepuesto:** especifica cómo se muestran las frases que se solapan
 - **ninguno (mantener todos):** no hay ningún filtro, se muestran las frases que se solapan.
 - **priorizar las frases más largas:** sólo se conserva la frase más larga y todas las demás se filtran.
 - **priorizar las frases más frecuentes:** se muestran las frases que aparezcan con mayor frecuencia en todo el documento.

Palabras vacías

Voyant Tools tiene una lista de palabras clave comunes que se han eliminado del corpus, como las palabras: “un”, “y”, “o”, “pero”, etc.²

Es posible eliminar palabras que no añaden mucho significado a nuestro análisis, para ello procedemos a filtrarlas. Por ejemplo, vamos a filtrar las palabras “assi”, “delos” y “otrossi” de la nube de palabras. Para ello primero colocamos el cursor sobre la barra gris en la parte superior de la ventana de la nube de palabras hasta que aparezca un menú de iconos. Luego hacemos clic en el icono azul de opciones . En el menú desplegable Opciones podemos modificar la configuración de la herramienta. Estas son las opciones disponibles:



- **Palabras excluidas:** permite definir un conjunto de palabras vacías que van a ser excluidas (la [guía de palabras vacías](#) ofrece más información)
- **Lista blanca:** permite definir un conjunto de palabras permitidas (lo opuesto a una lista de palabras vacías), solo los términos de esta lista se mostrarán en **Cirrus** (hay que tener en cuenta que la lista de palabras vacías aún está activa, por lo que debemos elegir “Ninguna” en el menú de palabras vacías para desactivarla)
- **Categorías:** permite especificar categorías en función de la frecuencia de las palabras.
- **Familia de fuentes:** determina qué fuente utiliza Cirrus. Aquí también se puede especificar una fuente instalada en nuestro ordenador, pero, por supuesto, es posible que no esté disponible en otros ordenadores (en cuyo caso se usa una fuente predeterminada, segura para la web)
- **Paleta:** permite editar la [paleta de colores](#).

² Estas son las palabras que aparecen en la lista de palabras vacías: *al, alguna, algunas, alguno, algunos, algún, ambos, ampleamos, ante, antes, aquel, aquellas, aquellos, aquí, arriba, atras, bajo, bastante, bien, cada, cierta, ciertas, ciertos, como, con, conseguimos, conseguir, consigo, consigue, consiguen, consigues, cual, cuando, de, del, dentro, donde, dos, el, ellas, ellos, empleais, emplean, emplear, empleas, empleo, en, encima, entonces, entre, era, eramos, eran, eras, eres, es, esta, estaba, estado, estais, estamos, estan, estoy, fin, fue, fueron, fui, fuimos, gueno, ha, hace, haceis, hacemos, hacen, hacer, haces, hago, incluso, intenta, intentais, intentamos, intentan, intentar, intentas, intento, ir, la, largo, las, lo, los, mientras, mio, modo, muchos, muy, nos, nosotros, o, otro, para, pero, podeis, podemos, poder, podria, podriais, podriamos, podrian, podrias, por, por qué, porque, primero desde, puede, pueden, puedo, que, quien, sabe, sabeis, sabemos, saben, saber, sabes, se, ser, si, siendo, sin, sobre, sois, solamente, solo, somos, soy, su, sus, también, teneis, tenemos, tener, tengo, tiempo, tiene, tienen, todo, trabaja, trabajais, trabajamos, trabajan, trabajar, trabajas, trabajo, tras, tuyo, ultimo, un, una, unas, uno, unos, usa, usais, usamos, usan, usar, usas, uso, va, vais, valor, vamos, van, vaya, verdad, verdadera cierto, verdadero, vosotras, vosotros, voy, y, yo*

Opciones

Palabras
excluidas:

Auto-detección

Editar lista

☒ Aplicar a todo

Lista blanca:

Editar Lista

Categorias:

auto

Editar

Familia de
fuentes:

Palatino

Paleta:

default

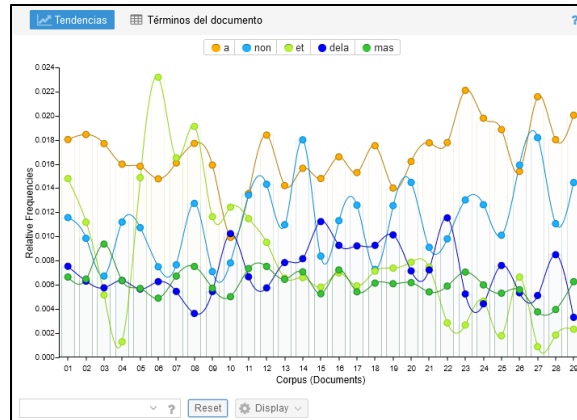
Lista de edición

↺ Reiniciar

✕ Cancelar

✓ Confirmar

[illegible]



Paso 4: Tendencias y frecuencia de las palabras

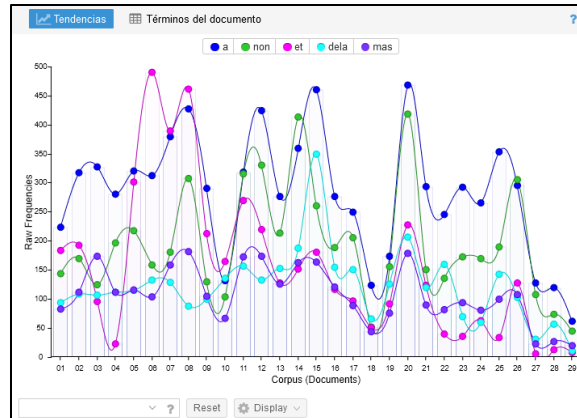
La herramienta **Tendencias** muestra un gráfico de líneas con las palabras más utilizadas en el corpus. Cada serie del gráfico está coloreada según la palabra que representa. En la parte superior del gráfico, una leyenda muestra qué palabras están asociadas con ciertos colores. Se puede hacer clic en las palabras de la leyenda para que se muestren o no. Al pasar el cursor sobre cualquier punto del gráfico, aparece un cuadro de llamada con información sobre el término seleccionado y su frecuencia.

De forma predeterminada, la herramienta **Tendencias** muestra las frecuencias relativas de las palabras en el corpus. Para ver el recuento absoluto de cada documento del corpus, debemos seleccionar **Frecuencias sin pulir**, que podemos encontrar al hacer clic en el icono azul de opciones en la barra de menú gris.

En la ventana emergente Opciones, hay que hacer clic en **Frecuencias sin pulir** y, a continuación, pulsar **Confirmar**. La herramienta **Tendencias** debería actualizarse automáticamente con el recuento absoluto de cada una de las palabras principales del corpus.

La ventana 'Opciones' permite configurar la visualización de los datos. Las opciones visibles son:

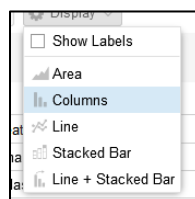
- Palabras excluidas:** keywords-130ae13591C (con botón 'Editar lista')
- Aplicar a todo:** ☒
- Categorías:** auto (con botón 'Edit')
- Segmentos:** 10 (con barra deslizante)
- Frecuencias:** ☒ Sin pulir, ☐ Relativo
- Paleta:** default (con botón 'Lista de edición')
- Botones de acción:** Reset, Cancelar, Confirmar



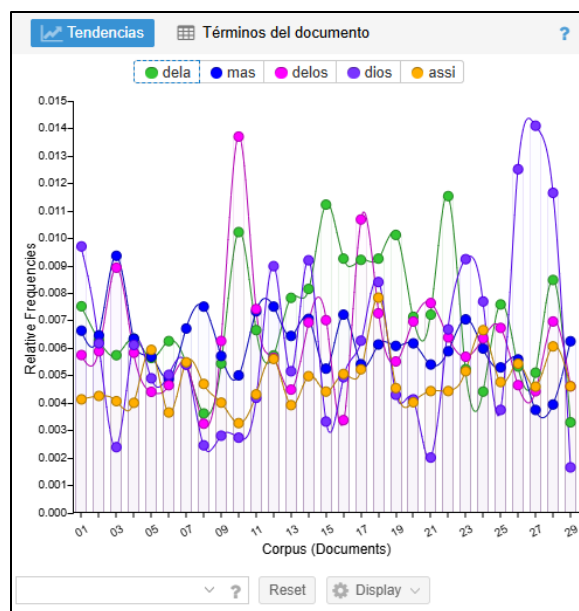
En cualquier momento podemos hacer clic en el botón **Reset** para volver a los valores predeterminados de la herramienta. El botón **Display** tiene dos componentes, **Show labels** y **Modo de gráfico**.

- Mostrar etiquetas: determina si cada elemento del gráfico debe etiquetarse
- Modo de gráfico:
 - **Área**: un gráfico de área (las etiquetas no están disponibles para este tipo de gráfico)
 - **Barras**: cada elemento es su propia columna para cada categoría
 - **Línea**: gráfico de líneas en todas las categorías
 - **Barra apilada**: gráfico de barras apiladas (los valores se muestran en columnas)
 - **Línea y barra apilada**: gráfico de líneas y barras apiladas superpuestas (el utilizado por defecto)

Para cambiar la visualización a barras debemos seleccionar el menú desplegable **Display** y, a continuación, haz clic en **Columns**.



El gráfico debería actualizarse automáticamente con un gráfico de barras.



y al hacer clic en ella se muestra la tabla de frecuencias de los términos en el documento con las siguientes columnas:

#	Términos	Contar	Relativo	Tendencia
<input type="checkbox"/>	2... rey	54	17,734	
<input type="checkbox"/>	2... moysen	103	15,585	
<input type="checkbox"/>	2... dios	83	14,096	
<input type="checkbox"/>	1... delos	181	13,702	
<input type="checkbox"/>	1... annos	174	13,172	
<input type="checkbox"/>	2... dios	240	12,512	
<input type="checkbox"/>	2... dios	77	11,651	
<input type="checkbox"/>	2... tierra	200	11,632	
<input type="checkbox"/>	2... dela	159	11,531	
<input type="checkbox"/>	1... dela	349	11,221	
<input type="checkbox"/>	1... moysen	78	11,103	
<input type="checkbox"/>	1... moysen	255	11,060	
<input type="checkbox"/>	2... moysen	145	10,972	
<input type="checkbox"/>	1... delos	174	10,679	
<input type="checkbox"/>	1... delas	167	10,250	
<input type="checkbox"/>	1... aaron	72	10,249	

- **#:** es el número de documento
- **Términos:** es la palabra en sí
- **Contar:** es la frecuencia con la que aparece la palabra
- **Relativo:** es la frecuencia relativa del término en el documento (se calcula dividiendo la frecuencia absoluta por el número total de términos del documento y multiplicando por 1 millón).
- **Tendencia:** es un gráfico que muestra la distribución del término en los segmentos del

documento; puedes pasar el ratón por encima del gráfico para ver resultados más precisos.

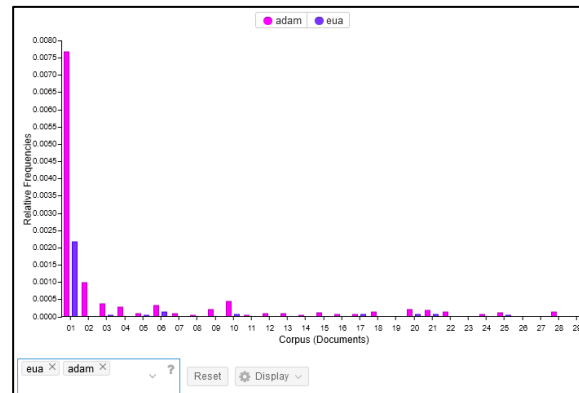
- **Nivel significativo:** mide la importancia de un término en relación con el resto de los términos del documento mediante la puntuación TF-IDF (esta columna no se muestra inicialmente)
- **Puntuación Z:** es la puntuación Z (puntuación estándar) de la frecuencia bruta del término en comparación con otras frecuencias de términos del mismo documento (esta columna no se muestra inicialmente, al añadir la columna de la puntuación Z puede ocultar la columna Tendencia)

Es posible gestionar las columnas pasando el ratón por encima de la barra de título de una columna y abriendo el menú desplegable, marcando allí las columnas que deseamos ver y las que no.

Paso 5: Búsquedas

Para ver un término en particular en el corpus, la herramienta **Tendencias** tiene un cuadro de búsqueda, donde se puede especificar una búsqueda avanzada, siendo posible buscar varias palabras al mismo tiempo.

Busquemos las palabras “adam” y “eua” juntas y veamos con qué frecuencia aparecen en el corpus. Para ello hay que escribir las palabras “adam” y “eua” como palabras separadas y, a continuación, hacer clic en **Intro**. La herramienta **Tendencias** se actualiza con el número de veces que estas palabras aparecen en el corpus. Hay que tener en cuenta que *Voyant Tools* no captura a *Adam* con una “A” mayúscula. Cada palabra se convierte a minúsculas, por lo que es preciso usar “adam” y “eua” en su lugar.



Como se puede ver al examinar el gráfico de tendencias anterior, “adam” aparece 174 veces y “eua” aparece 38 veces. El cuadro de búsqueda de todas las herramientas permite realizar diferentes consultas avanzadas, siendo estos algunos ejemplos:

- eua: coincide con el término exacto “eua”
- eua*: hace coincidir todos los términos que comienzan con el prefijo “eua-” y luego un comodín como un solo término
- ^eua*: hace coincidir los términos que comienzan con “eua-” como términos separados (eua, euas, euad, euangelio, euandro, etc.)
- *eron: coincidir con los términos que terminan con el sufijo “-eron” como un solo término

- ^*eron: hace coincidir los términos que terminan con el sufijo “-eron” como términos separados (dieron, quisieron, oyeron, touieron, etc.)
- eua, adam: empareja cada término separado por comas como términos separados
- eua|adam: hace coincidir los términos separados por barras verticales como un solo término
- “dios fizo”: como una frase exacta que respeta el orden de las palabras
- “dios fizo”~0: empareja las combinaciones *dios fizo* y *fizo dios* como una frase en la que no importa el orden de las palabras, separadas por 0 palabras
- “adam eua”~5: Empareja a “adam” cerca de “eua”, en cualquier orden, separados por no más de 5 palabras

También es posible utilizar [] indicando que cualquiera de los caracteres dentro de los corchetes puede aparecer como mucho una vez, así *ali[nm]p[ijy]** y *li[nm]p[ijy]** sirven para encontrar las variantes ortográficas de *alimpiar* y *limpiar*.

Paso 6: Palabras en contexto

Además del examen de las frecuencias de las palabras, *Voyant Tools* permite también examinar las palabras en su contexto.

Herramienta Contextos

La herramienta **Contextos** (o Palabras clave en contexto) muestra cada aparición de una palabra clave con un poco de texto circundante (el contexto). La vista de tabla muestra las tres siguientes columnas:

- Documento: muestra el documento en el que aparece la palabra clave
- Izquierda: muestra las palabras contextuales a la izquierda de la palabra clave
- Término: muestra la palabra clave que coincide con la consulta
- Derecha: muestra las palabras contextuales a la derecha de la palabra clave

De forma predeterminada, los contextos se muestran para las palabras más frecuentes del corpus. Para buscar una palabra diferente, es necesario especificar un término en el cuadro de búsqueda de la herramienta **Contextos**.

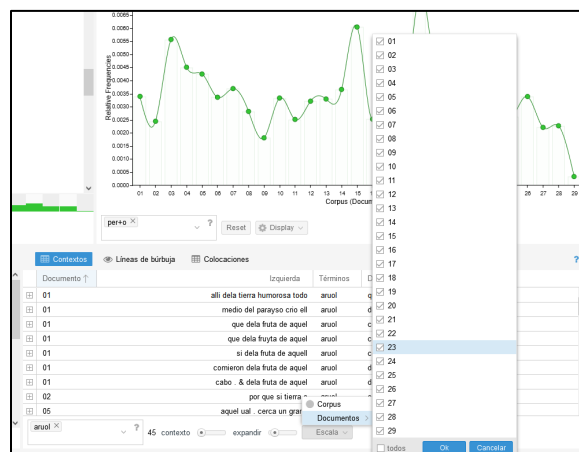
Examinemos ahora la frecuencia de la palabra “aruol” en su contexto. Para ello escribimos “aruol” en el cuadro de búsqueda y hacemos clic en **Intro**. El término “aruol” aparecerá con las palabras que lo rodean (el contexto), organizadas por libro.


The screenshot shows the 'Contextos' tool interface with the search term 'aruol' entered in the search box. The results are displayed in a table with four columns: Documento, Izquierda, Términos, and Derecha. The results are organized by document (01, 02, 05).

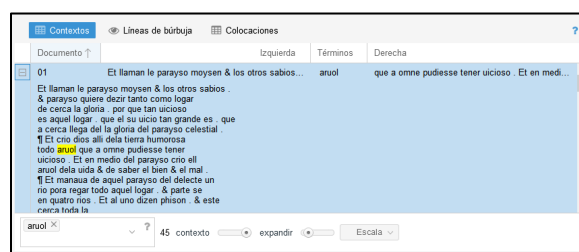
Documento	Izquierda	Términos	Derecha
01	ali dela tierra humerosa todo	aruol	que a omne pudiesse tener
01	medio del parayso crio ali	aruol	dela vida & de saber el
01	que dela fruta de aquel	aruol	comiesse que non era mortal
01	que dela fruta de aquel	aruol	comiesse que se tomase mortal
01	si dela fruta de aquel	aruol	comiesse . que muerre mome lascas
01	contienor dela fruta de aquel	aruol	de medio del parayso de
01	cabo . & dela fruta de aquel	aruol	de saber el bien & el
02	por que si tierra o	aruol	o alguna cosa fallasse descubierta
05	aquel ual . cerca un grand	aruol	que aue y estonces . & este

At the bottom of the interface, there is a search box containing 'aruol', a frequency count of '45 contextos', and buttons for 'expandir' and 'Escala'.

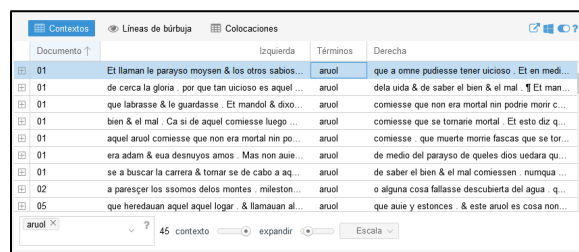
De forma predeterminada, se muestran todas las formas de la palabra en el corpus. Si solo se quieren ver los resultados de un solo libro, hay que hacer clic en el menú desplegable **Escala** y seleccionar **Documentos**. Esto generará una lista con los documentos en el corpus. A continuación, se debe marcar el libro que se desea ver en contexto.

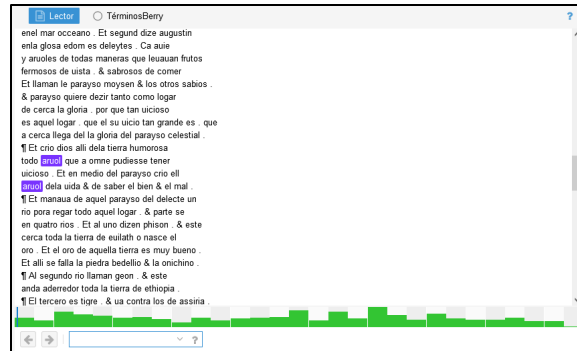


También se puede aumentar el contexto que rodea a la palabra consultada, para ello hay que utilizar el control deslizante **Contexto**; el valor predeterminado es 5 y el máximo 50. Del mismo modo, el control deslizante **Expandir** determina cuántas palabras se muestran cuando se expande una fila determinada (haciendo clic en el icono  en la columna más a la izquierda); el valor predeterminado es 50, el mínimo es 5 y el máximo es 500 palabras.



Cuando seleccionamos una fila concreta en la herramienta **Contextos**, la herramienta **Lector** se actualizará con el término de búsqueda resaltado en amarillo, mostrando también dónde aparece esa fila de texto en el corpus.





Herramienta Lector

La herramienta **Lector** muestra todo el corpus y se compone de dos componentes visuales: el lector de texto y el visor. El lector de texto muestra todo el texto del corpus. Con el lector de texto es posible:







- Avanzar dentro del lector de texto para obtener más contenido
- Colocar el cursor sobre una palabra para mostrar su frecuencia en el documento
- Hacer clic en una palabra para buscarla en el **Lector** (y otras herramientas si corresponde)

El visor muestra una descripción general de todo el corpus. Esto es especialmente útil cuando hay varios documentos en un corpus. Las barras representan cada documento en el orden en que aparecen en el corpus.



La longitud relativa del documento se representa tanto vertical como horizontalmente (en otras palabras, cuanto más alto y ancho se muestra un documento, más largo es). La barra vertical azul indica la posición actual del lector de texto en el corpus. Se puede hacer clic en cualquier lugar a lo largo del visor para saltar a otra ubicación. Para avanzar o retroceder en el texto, deben usarse las flechas junto al cuadro de búsqueda.

Paso 7: Otras herramientas







Además de las cinco herramientas que aparecen de forma predeterminada en la interfaz de trabajo (**Cirrus**, **Lector**, **Tendencias**, **Sumario** y **Contextos**), es posible acceder a otra serie de herramientas. Para ello es necesario colocar el cursor sobre el símbolo del signo de interrogación  en la barra gris en la parte superior de cada una de las herramientas hasta que aparezca el menú de iconos    , hacer clic en el icono de la ventana  y seleccionar la herramienta con la que se desea trabajar. Estas son algunas de las herramientas disponibles:

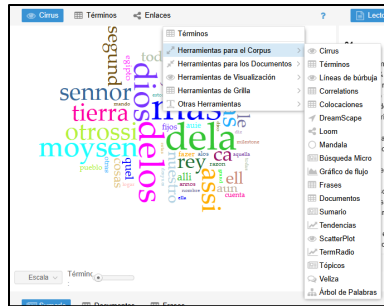
- **Arco textual** (TextualArc): es una visualización de los términos de un documento que incluye un centroide ponderado de términos y un arco que sigue los términos en el orden del documento.
- **Colocaciones** (Collocates): es una tabla que muestra los términos que aparecen con más frecuencia cerca de palabras clave en todo el corpus.

- **Correlaciones** (Correlations): permite explorar en qué medida las frecuencias de los términos varían en sincronía (términos cuyas frecuencias aumentan y disminuyen juntas o inversamente).
- **Documentos** (Documents): es una tabla de los documentos del corpus e incluye funciones para modificar el corpus.
- **Enlaces** (Links): representa las palabras clave y los términos que aparecen en estrecha proximidad como un grafo de red.
- **Frases** (Phrases): muestra secuencias repetidas de palabras organizadas por frecuencia de repetición o número de palabras en cada frase repetida.
- **Gráfico de dispersión** (ScatterPlot) es una visualización gráfica de cómo se agrupan las palabras en un corpus utilizando la similitud de documentos, el análisis de correspondencias o el análisis de componentes principales.
- **Líneas de burbuja** (Bubblelines): visualiza la frecuencia y la distribución de los términos en un corpus.
- **Micro búsqueda** (Microsearch): es una visualización de la frecuencia y distribución de términos en cada uno de los documentos que conforman un corpus.
- **StreamGraph**: es una visualización que muestra el cambio de la frecuencia de las palabras en un corpus (o dentro de un único documento).
- **Temas** (Topics): proporciona una forma rudimentaria de generar grupos de temas a partir de un documento o corpus y ver después cómo se distribuye cada tema (grupo de términos) en el documento o corpus.
- **Términos del corpus** (Corpus Terms): es una tabla de frecuencias de términos en todo el corpus.
- **Términos del documento** (Document Terms): es una tabla de las frecuencias de los términos de uno de los documentos del corpus.

Para obtener más información sobre estas herramientas y sus funcionalidades, visite la [guía de ayuda](#) de *Voyant Tools*.

Trabajando con las colocaciones

Mientras que la herramienta de **Contextos** permite ver qué palabras rodean una palabra clave en el corpus, la herramienta **Colocaciones** muestra qué términos aparecen con más frecuencia cerca unos de otros. Echemos un vistazo a los términos que aparecen muy cerca de la palabra “adam”. Para ello debemos colocar el cursor sobre el símbolo del signo de interrogación  en la barra gris en la parte superior de cada una de las herramientas hasta que aparezca el menú de iconos    . A continuación hacemos clic en el icono de la ventana , y en *Herramientas para el Corpus* seleccionamos **Colocaciones**.



Una vez seleccionada, la herramienta **Colocaciones** debería aparecer automáticamente y reemplazar la herramienta **Cirrus**.

Término	Contexto del término	Contar (contexto)
delos	delos	716
sennor	dios	639
dela	dela	618
delas	delos	515
delos	delas	511
delas	delas	500
dios	nuestro	446
delos	dela	432
dela	delos	425
dela	tierra	418
sennor	moysen	414
moysen	sennor	410
mas	mas	399
moysen	nuestro	389
delas	cosas	330
delos	otros	316

Ahora escribimos la palabra “adam” en el cuadro de búsqueda vacío de la herramienta **Colocaciones** y al hacer clic en **Enter** veremos una de tabla con las palabras más frecuentes que aparecen muy cerca de la palabra “adam”.

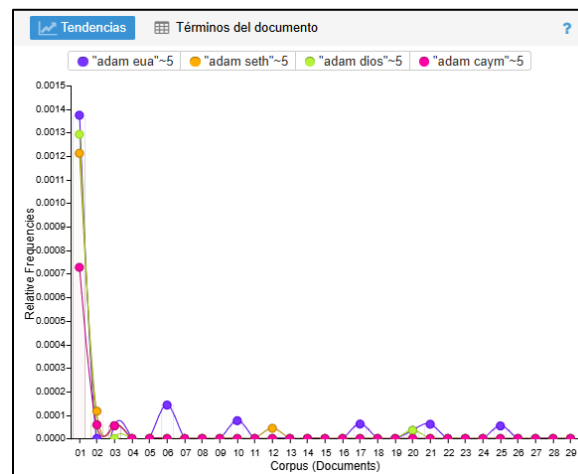
Término	Contexto del término	Contar (contexto)
adam	eua	25
adam	annos	21
adam	fasta	18
adam	seth	17
adam	mundo	17
adam	fecho	17
adam	mill	16
adam	dios	16
adam	fijos	14
adam	fizo	12
adam	fijo	12
adam	noe	11
adam	comienço	11
adam	caym	10
adam	otros	9
adam	otras	6

De forma predeterminada, la vista de tabla muestra las tres siguientes columnas:

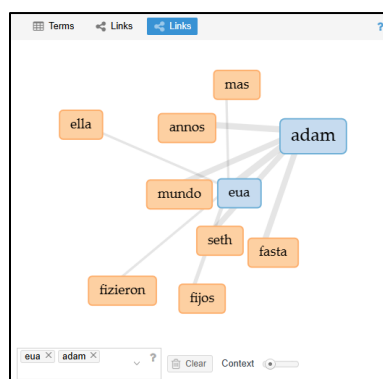
- **Término:** muestra la(s) palabra(s) clave que se están buscando.
- **Colocaciones:** son las palabras que se encuentran en la proximidad de cada palabra(s) clave(s).
- **Contar (contexto):** muestra el recuento de frecuencia de la colocación que se produce en las proximidades de la(s) palabra(s) clave.

La barra deslizante **contexto** determina el número de palabras a ambos lados del término buscado. El valor predeterminado es 5 palabras a cada lado, pero puede cambiarse entre 1 y 30 palabras.

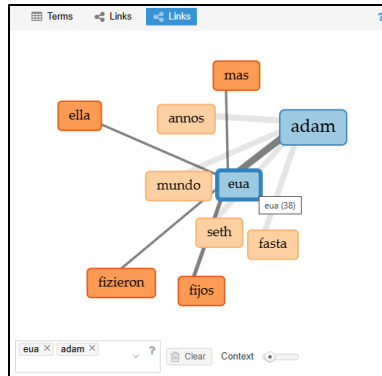
Como podemos ver, la palabra “adam” aparece con mayor frecuencia (25 veces para ser exactos) con el colocado “eua”. Las variaciones de esta colocación aparecen en el corpus, con “adam” y “annos” también ocurriendo 21 veces, y “adam” y “fasta” ocurriendo 18 veces. Para comparar las colocaciones en el corpus, es preciso marcar las colocaciones que deseamos ver para que la herramienta **Tendencias** se actualice automáticamente. Comparemos ahora las colocaciones de “adam” con “eua”, “dios”, “caym” y “seth”.



La herramienta **Enlaces** permite analizar las colocaciones de una forma más visual. Para acceder a ella podemos hacer clic en **Enlaces** y escribir las palabras “adam” y “eua” en el cuadro de búsqueda vacío de la herramienta **Enlaces**.



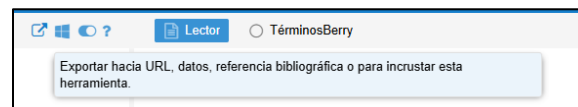
Al pasar el ratón sobre uno de los términos se muestra su frecuencia absoluta (38 en el caso de “eua”) y los enlaces con otros términos cambian de color mostrando las conexiones.




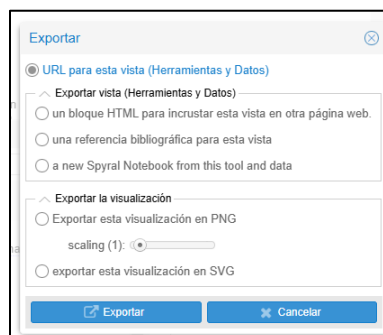
La barra deslizante `contexto` determina el número de palabras a ambos lados del término buscado. El valor predeterminado es 5 palabras a cada lado, pero puede cambiarse entre 1 y 30 palabras.

Paso 8: Cómo incorporar *Voyant Tools* en otros sitios web

Voyant Tools está diseñado para funcionar como un entorno independiente (voyant-tools.org) o como un conjunto de módulos independientes que pueden ser integrados en otros sitios web. Para obtener un enlace debemos utilizar la función de exportación colocando el cursor sobre la barra gris hasta que aparezca un menú de iconos:




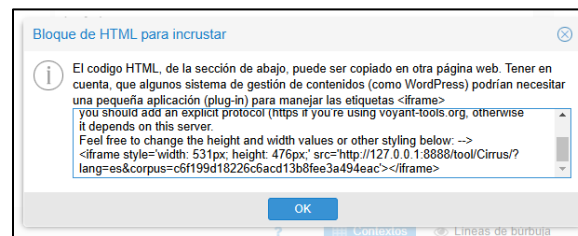
Al hacer clic en el icono Exportar  aparece un menú con las siguientes opciones para exportar.



- URL para esta vista (Herramientas y Datos)
 - Genera una URL para la herramienta y los datos visualizados en ese panel; es la opción predeterminada.
- Exportar Vista (herramientas y datos)
 - un bloque HTML para incrustar esta vista en otra página web: genera un fragmento de código HTML, que se puede utilizar para insertar esta sesión de *Voyant Tools* en una página web.

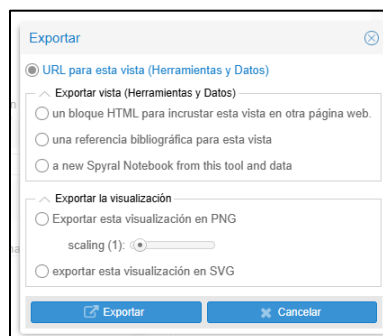
- una referencia bibliográfica para esta vista: genera la referencia bibliográfica para esta sesión de *Voyant Tools*.
- a new Spyral Notebook from this tool and data: Spyral Notebook es un entorno de programación construido sobre *Voyant Tools* y que lo amplía. Para más información sobre Spyral Notebook puede consultarse la [guía de ayuda](#).
- Exportar la visualización:
 - exportar esta visualización en PNG: genera una imagen en formato PNG de la herramienta y un fragmento de código HTML que contiene la imagen.
 - exportar esta visualización en SVG: genera una imagen en formato SVG de la herramienta y un fragmento de código HTML que contiene la imagen.

Vamos ahora a generar un bloque HTML en la herramienta **Cirrus**. Para ello hacemos clic en el icono Exportar  y seleccionamos **Exportar vista (Herramientas y Datos): un bloque HTML**. Debería aparecer una ventana emergente con un fragmento de código HTML (<iframe >) que puede utilizarse para insertar la visualización en otro sitio web.



Paso 9: Cómo exportar imágenes desde *Voyant Tools*

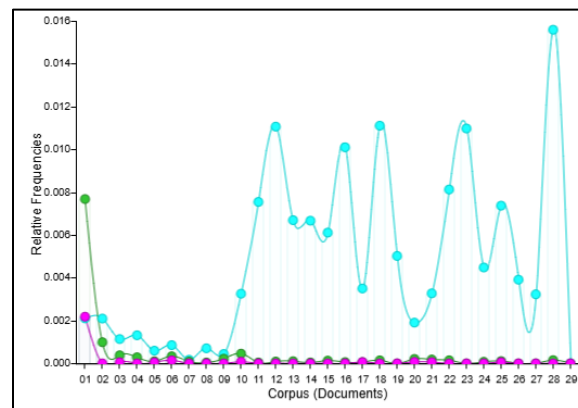
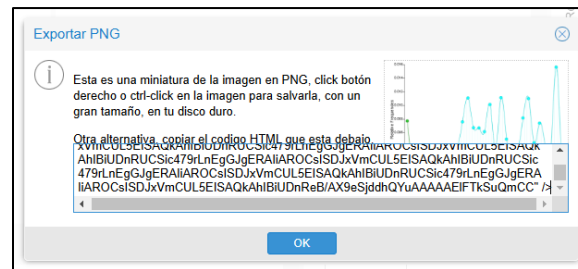
Si se desea exportar una imagen de *Voyant Tools*, hay que seleccionar **exportar esta visualización en PNG** en “Exportar visualización”.



“Exportar visualización”, disponible en alguna de las herramientas, genera una imagen en formato PNG de la herramienta actual y un fragmento de código HTML que contiene la imagen. Una vez seleccionada esta opción, es posible cambiar la escala de la imagen usando la barra deslizante de escala, que aumentará o disminuirá el tamaño de la imagen.

Vamos ahora a exportar el gráfico en la herramienta **Tendencias** mostrando la distribución de las

palabras “adam”, “eua”, y “moysen” en GE1. Para ello debemos introducir esas palabras en el cuadro de búsqueda y a continuación hacemos clic en el icono Exportar y seleccionamos **Exportar esta visualización en PNG**. Debería aparecer una ventana emergente con una miniatura de la imagen³ (podemos abrirla en otra ventana del navegador) y un fragmento de código HTML que podemos utilizar para insertar la imagen de la herramienta y el corpus seleccionados en otro sitio web.



Paso 10: Recursos adicionales para *Voyant Tools*

Este tutorial sólo ofrece una visión superficial de lo que puede hacer con *Voyant Tools* y los tipos de análisis de texto que puede realizar. Para obtener más información sobre *Voyant Tools*, consulta los siguientes recursos:

- [Voyant Help Guide](#)
- [Voyant Tools GitHub](#)
- [Voyant Tools Twitter: @VoyantTools](#)

³ Desafortunadamente la leyenda con los términos consultados no aparece incluida en la imagen. Tenemos entonces la opción de exportar los datos actuales en formato TSV para su posterior procesamiento en Excel o programas similares.

Ficheros:

<https://hispanicseminary.org/UIMP2024/GE1.zip>

<https://hispanicseminary.org/UIMP2024/poesia.zip>

Enlaces:

GE1 (texto completo): <https://voyant-tools.org/?corpus=79ff6f5748e49400315b122a0d7a75b8>

GE1 (29 libros): <https://voyant-tools.org/?corpus=92f735019a4381541c418064a5773d59>

GE1 (986 capítulos): <https://voyant-tools.org/?corpus=4253621ee61a7cfd1a95efb188ba0f81>

Poesía (10 obras): <https://voyant-tools.org/?corpus=587a983b1c8162cd00a3c3af41206301>