

ALGUNAS IDEAS A PARTIR DE UNA EXPERIENCIA

# Una biblioteca que ayuda a investigar un acervo

Preguntas y decisiones para  
preparar aplicaciones de IA  
en centros de documentación  
desde la perspectiva de un  
antropólogo

Dr Alejandro Flores-Aguilar  
AVANCSO



Open Knowledge

# GLOSARIO

**AVANCSO:** la Asociación para el Avance de las Ciencias Sociales en Guatemala (AVANCSO) es un centro de investigación en ciencias sociales que inició su actividad en enero de 1986, luego de producirse la apertura política que condujo a la instauración en Guatemala de un gobierno civil electo.

**BÚSQUEDA SEMÁNTICA:** es una tecnología de motor de búsqueda que interpreta el significado de las palabras y las frases. Los resultados de una búsqueda semántica devolverán contenido que coincida con el significado de una búsqueda, en lugar de contenido que coincida literalmente con las palabras en la búsqueda.

**CAJA DE ARENA:** un arenero, caja de arena o *sandbox* es un entorno de prueba aislado que permite a los usuarios ejecutar programas o archivos sin afectar la aplicación, el sistema o la plataforma en la que se ejecutan.

**CHATBOT:** es un programa informático diseñado para simular una conversación con personas a través de texto o voz.

**COINCIDENCIA TEXTUAL:** también conocida como búsqueda exacta o búsqueda por palabra clave, es un método de recuperación de información donde el sistema localiza cadenas de caracteres que coinciden de forma idéntica con el texto ingresado por el usuario.

**GENERACIÓN AUMENTADA POR RECUPERACIÓN (RAG):** es una técnica de inteligencia artificial que optimiza las respuestas de un modelo de lenguaje (como GPT o Gemini) al conectarlo con una base de conocimientos externa o privada. En lugar de depender únicamente de sus datos de entrenamiento, el sistema busca fragmentos de información relevantes en documentos específicos (como PDF, manuales o bases de datos) y se los entrega al modelo para que genere una respuesta precisa basada estrictamente en esa fuente real.

**JSONL (JSON Lines):** el formato JSONL guarda cada documento o texto en una línea de código independiente. Es el estándar preferido en la Inteligencia Artificial porque permite a los modelos leer y procesar bibliotecas masivas de datos de forma rápida, eficiente y sin saturar la memoria de la computadora.

**METADATOS:** son un conjunto de datos que describen, explican o dan contexto a otros datos. Se les conoce comúnmente como “datos sobre los datos”. Esta información puede indicar quién creó un archivo, cuándo se modificó, de qué trata, en qué formato está guardado o cuál es su tamaño. Los metadatos facilitan la clasificación y localización rápida de documentos, imágenes o archivos, optimizando los procesos de búsqueda interna dentro de las organizaciones y sistemas informáticos.

**README:** es un documento informativo que acompaña a un proyecto de software, una aplicación o un repositorio para explicar su propósito, contenido, historia y uso. Es usualmente el primer archivo que un desarrollador de un proyecto lee.

**RECONOCIMIENTO ÓPTICO DE CARACTERES (OCR):** es el proceso por el cual se convierte una imagen de texto en un formato de texto que pueden leer las máquinas. Por ejemplo, si se escanea un formulario, la computadora guarda el escaneo como un archivo de imagen donde no se puede utilizar un editor de texto para editar, buscar o contar las palabras del archivo de imagen. Sin embargo, se puede utilizar el OCR para convertir esa imagen en un documento de texto con su contenido como datos de texto.

**SQLite:** sistema que guarda bases de datos completas en un solo archivo dentro de la computadora. Mientras que JSONL es un formato de texto pensado para leer documentos de corrido de principio a fin, SQLite funciona como un archivador inteligente que permite buscar, filtrar y organizar información específica al instante antes de procesarla.

# Para qué sirven estas páginas

Estas páginas están dirigidas a quienes trabajan en centros de documentación (desde ahora en adelante “centros”) y se preguntan qué hace falta para preparar un acervo destinado a aplicaciones de inteligencia artificial (IA). El texto ofrece un punto de partida para organizar las primeras decisiones de un piloto. La pregunta orientadora es: ¿qué tendría que hacer un centro para que una aplicación ayude a investigar sus materiales y permita volver a los documentos y las páginas que sostienen sus resultados?

El punto de partida de esta guía es mi experiencia como antropólogo construyendo una biblioteca digital de mi propia colección con un modelo local de IA que funciona en mi computadora. Durante años reuní libros, artículos, capítulos, entrevistas y otros documentos que fueron quedando almacenados en mi computadora. A medida que la colección crecía, la forma en que había organizado los archivos hasta entonces se quedó corta. Los problemas que fui afrontando eran bastante sencillos: a veces recordaba el título de un texto, pero el nombre del archivo no tenía ninguna relación evidente con él y el motor de búsqueda integrado en mi sistema operativo tampoco lograba encontrarlo. Podía pasar bastante tiempo buscando un documento que sabía que estaba en la computadora. Esas interrupciones fueron volviendo más lentos mis procesos de investigación.

Como ya corría un modelo local de IA, pensé en construir una pequeña aplicación que leyera los PDF y les asignara nombres más útiles. Funcionó mal. Para proponer un nombre, el programa tenía que reconocer qué documento tenía delante. Los metadatos<sup>1</sup> podían estar ausentes o conservar información equivocada. Cuando el PDF provenía de un escaneo, el programa a veces ni siquiera podía leer el título.

Ese primer experimento cambió la pregunta que me estaba planteando. El programa tenía que reconocer primero qué documento contenía cada archivo. Si las páginas sólo existían como imágenes, también tenía que convertirlas en texto. Después, cada página debía conservar su relación con el documento original. Quizá esa misma infraestructura me permitiría encontrar pasajes específicos de

<sup>1</sup> Los metadatos describen el contenido, el contexto y la procedencia de los datos de manera estandarizada y estructurada. Véase el apartado ‘Glosario’.

artículos que había leído muchos años antes o seguir la trayectoria de un concepto a través de una literatura más amplia.

Así comencé a construir lo que podía llamarse un *chatbot*<sup>2</sup> bibliográfico. La expresión era atractiva, aunque pronto resultó engañosa. La ventana de conversación concentraba la atención. Sin embargo, buena parte del trabajo ocurría antes de que alguien escribiera una pregunta.

Fui construyendo dos formas de búsqueda. Una era determinista y trabajaba con títulos, autores, fechas y coincidencias exactas. La otra intentaba recuperar relaciones entre materiales, aunque no compartieran las mismas palabras. Durante meses, la biblioteca fue más bien un conjunto de carpetas, páginas extraídas y tablas que intentaban recordar de qué documento había salido cada cosa.

Después de conversar con la directora ejecutiva de Open Knowledge Foundation, empecé a considerar la posibilidad de llevar estas primeras experiencias al Centro de Documentación de la Asociación para el Avance de las Ciencias Sociales en Guatemala, AVANCSO. Mi colección podía funcionar como una “caja de arena”<sup>3</sup>, un espacio que toleraba mis decisiones provisionales y mis carpetas mal nombradas mientras aprendía.

Pero AVANCSO desplazó la pregunta: un centro de documentación requería otra clase de atención. El acervo de AVANCSO tiene otra clase de memoria. Sus materiales fueron reunidos durante décadas y distintas personas participaron en su cuidado. Parte del conocimiento sobre las relaciones entre los documentos permanece en las prácticas de quienes han trabajado con ellos. Una aplicación institucional tendría que aprender a trabajar con esa memoria distribuida.

El intercambio con Open Knowledge Foundation permitió pensar que algunos de estos aprendizajes podrían ser útiles para otras organizaciones. Una herramienta construida desde mi biblioteca ofrece un punto de partida. Sus categorías tendrían que discutirse al encontrarse con otro acervo. Cada centro necesitará decidir qué problema quiere atender; esa decisión permitirá establecer qué relaciones debe conservar y qué aplicación podría desarrollar desde sus propias condiciones. La pregunta comenzaba así a compartirse.

<sup>2</sup> Programa informático diseñado para simular una conversación con personas a través de texto o voz.

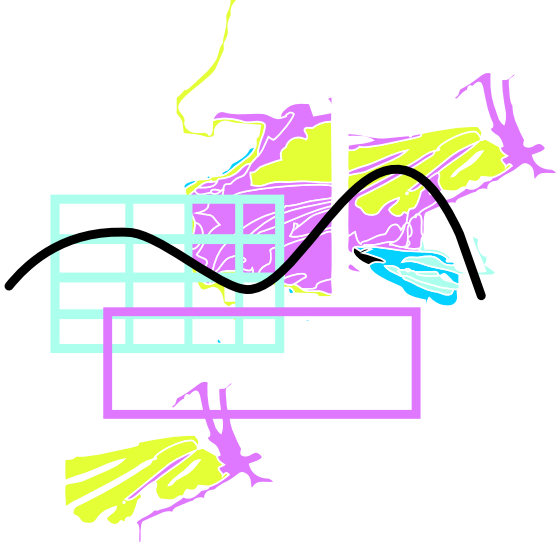
<sup>3</sup> Un entorno de prueba aislado que permite a los usuarios ejecutar programas o ejecutar archivos sin afectar la aplicación, el sistema o la plataforma en la que se ejecutan.

# Rutas de lectura

El orden de las secciones de este texto sigue el paso de mi biblioteca personal hacia la posibilidad de un piloto institucional. El índice permite seguir ese recorrido completo. Pero un centro también puede comenzar por el problema que necesita discutir. El siguiente diagrama puede ayudarlos:

|                                                                                                            |                                                                |    |                                                                                                                                         |
|------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------|----|-----------------------------------------------------------------------------------------------------------------------------------------|
| Si buena parte del acervo permanece en el papel: aquí se examinan las decisiones previas al procesamiento  | Para qué sirven estas páginas                                  | 4  | Si el centro ya cuenta con materiales digitalizados: aquí se ocupan de la trazabilidad, la recuperación y las respuestas generadas      |
|                                                                                                            | De una biblioteca personal a una pregunta compartida           | 7  |                                                                                                                                         |
|                                                                                                            | Comenzar por una consulta                                      | 9  |                                                                                                                                         |
|                                                                                                            | Documentos con historias y condiciones distintas               | 10 |                                                                                                                                         |
|                                                                                                            | Dónde circulan los documentos y quién puede decidir sobre ello | 11 |                                                                                                                                         |
|                                                                                                            | Soberanía como margen de decisión                              | 12 |                                                                                                                                         |
| Si la conversación se refiere a las aplicaciones que podrían construirse: aquí se desarrolla ese horizonte | Los documentos todavía tienen que entrar                       | 14 | Si el centro quiere delimitar un piloto y sostenerlo con el tiempo: aquí se reúnen las decisiones institucionales que tendrá que asumir |
|                                                                                                            | Una base que permita reconstruir el recorrido                  | 15 |                                                                                                                                         |
|                                                                                                            | Formas de buscar dentro de una biblioteca                      | 17 |                                                                                                                                         |
|                                                                                                            | Cuando la biblioteca empieza a responder                       | 18 |                                                                                                                                         |
|                                                                                                            | Una biblioteca que recuerda cómo se la consulta                | 20 |                                                                                                                                         |
|                                                                                                            | Un asistente que todavía lee                                   | 21 |                                                                                                                                         |
|                                                                                                            | Una mesa de trabajo alrededor del acervo                       | 22 |                                                                                                                                         |
|                                                                                                            | Los README como diario de campo                                | 25 |                                                                                                                                         |
|                                                                                                            | Preguntas para poner en marcha un piloto                       | 26 |                                                                                                                                         |
|                                                                                                            | Una infraestructura que sale de mi computadora                 | 28 |                                                                                                                                         |
| Una guía de pasos final                                                                                    | 29                                                             |    |                                                                                                                                         |

# De una biblioteca personal a una pregunta compartida



El experimento que hice para consultar mi biblioteca permite encontrar documentos y buscar dentro de sus páginas. También puede recuperar pasajes por su cercanía semántica y generar respuestas a partir de ellos. La base conserva el camino hacia las fuentes y algunas pruebas automáticas ayudan a revisar lo que ocurre cuando una respuesta pierde ese vínculo.

¿Cómo lo construí? Adelanto en estas páginas ese recorrido para brindar un panorama general.

Mi punto de partida —pero también el de un centro— es una consulta reciente que se quiera atender. Por ejemplo, una persona puede recordar un libro y no encontrarlo. También puede tener el documento a la vista y necesitar la página donde aparece un pasaje. Trabajar sobre una dificultad reconocible permite precisar qué función tendría la aplicación y qué necesitaría comprobar la persona antes de utilizar lo que encontró. La consulta elegida permite delimitar qué parte del acervo haría falta procesar.

El centro tendría que revisar las condiciones de acceso de esos documentos y decidir por cual infraestructura pueden circular. Si los materiales se encuentran en papel, por ejemplo, el escaneo producirá imágenes que el Reconocimiento Óptico de Caracteres (OCR, en sus siglas en inglés)<sup>4</sup> deberá convertir en texto. La base documental tendrá que conservar la relación entre ese texto, la página y el archivo original.

Cuando esas relaciones quedan registradas, la forma de buscar puede responder al tipo de pregunta. Una coincidencia textual<sup>5</sup> resulta útil para localizar un apellido o una frase. La búsqueda semántica<sup>6</sup> permite acercar pasajes que tratan un problema con vocabularios

4 El OCR es un programa que extrae y reutiliza datos de documentos escaneados, imágenes de cámara y PDF de sólo imagen. Véase el apartado 'Glosario'.

5 Un método donde el sistema busca las palabras exactas que escribes dentro de un documento o en internet. Véase el apartado 'Glosario'

6 Un método avanzado que busca por el significado y la intención detrás de tus palabras, en lugar de solo emparejar letras exactas. Véase el apartado 'Glosario'.

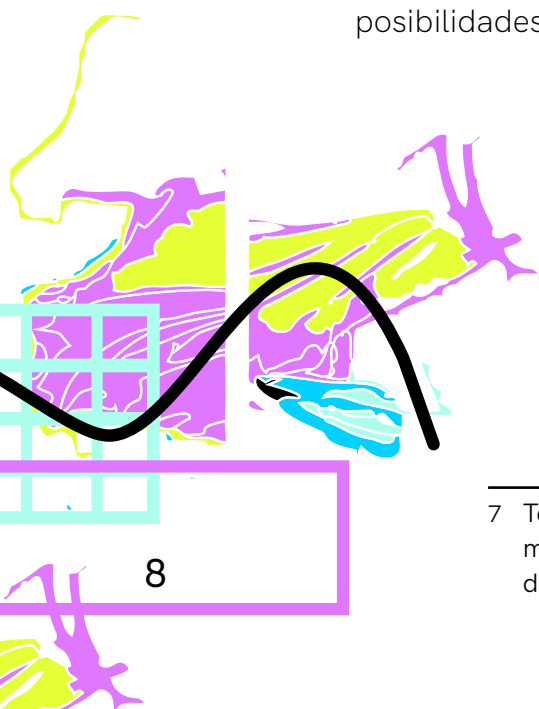
diferentes. En algunas consultas, los pasajes recuperados pueden entregarse a un modelo para que redacte una respuesta a partir de ellos. Esa operación se llama Generación Aumentada por Recuperación (RAG)<sup>7</sup>. Una etapa posterior podría permitir que la aplicación conservara un recorrido de investigación y lo retomara varios días después.

Hasta ahí llega la infraestructura que he construido.

Como veremos más adelante en este documento, cuestiones como el “asistente librero” pertenecen a la parte proyectiva del piloto de AVANCSO: la mesa de trabajo y la constelación de herramientas permiten imaginar cómo podría acompañar una investigación más prolongada. La infraestructura institucional para AVANCSO también tendrá que diseñarse y ponerse a prueba.

Como adelanté, AVANCSO cambia la escala de la pregunta comparada a mi biblioteca personal. Su centro de documentación conserva materiales reunidos durante décadas y las personas que lo consulten llegarán con conocimientos distintos de los míos. Las condiciones de acceso tendrán que decidirse junto con quienes conocen la historia del acervo y tienen autoridad sobre sus documentos. El piloto también deberá quedar en manos de personas que puedan continuarlo cuando cambien las herramientas o se retire quien construyó la primera versión.

Llamo heurístico al conjunto provisional de preguntas reunidas en estas páginas. Puede utilizarse para conversar con otros centros que estén considerando un trabajo semejante y para reconocer decisiones que suelen permanecer ocultas dentro de una aplicación. La discusión tendría que ocurrir junto a documentos concretos y personas responsables de cuidarlos. Cada proyecto deberá calcular el costo de preparar los materiales y establecer qué infraestructura puede mantener. Con esas respuestas podrá definir un piloto acorde con sus posibilidades.



<sup>7</sup> Técnica de Inteligencia Artificial que optimiza las respuestas de un modelo de lenguaje (como GPT o Gemini) al conectarlo con una base de conocimientos externa o privada. Véase el apartado ‘Glosario’.

## PUNTO DE PARTIDA

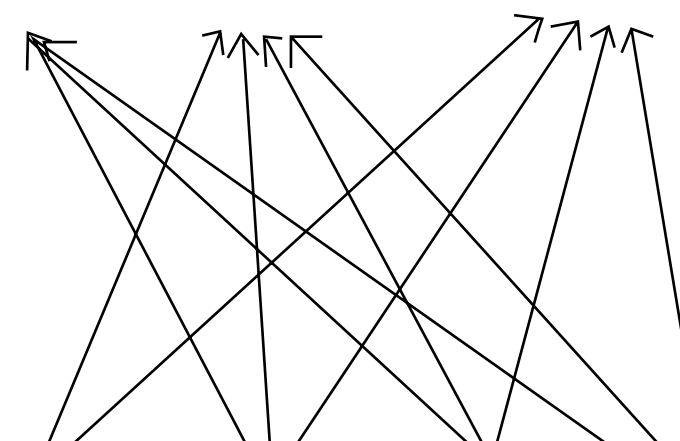
# Comenzar por una consulta

Mi biblioteca tenía una ventaja difícil de trasladar a un centro de documentación. Yo podía reconocer un título mal escrito o advertir que una respuesta no correspondía a lo que recordaba de una obra. Esa memoria personal compensaba algunas debilidades de la aplicación. Quienes consultaran el acervo de AVANCSO llegarían con referencias propias y necesitarían orientación aunque no supieran de antemano qué documento buscar.

Esta diferencia me llevó a pensar en la figura de un buen librero. Quien conoce una biblioteca desde hace tiempo puede escuchar una pregunta todavía imprecisa y orientar a alguien hacia documentos que quizá no había considerado. El librero sabe también cuándo conviene regresar al catálogo, abrir una caja o buscar una página antes de aventurar una respuesta.

La figura del librero sugería una manera de comenzar el diseño de la aplicación. Antes de escoger un modelo de IA, yo acompañaría algunas búsquedas. Querría saber qué está intentando hacer la persona y en qué punto se detiene. Una dificultad concreta permite comenzar por algo que el centro ya reconoce y puede describir.

Al acompañar esas consultas aparecerían necesidades distintas. Una persona podría recordar haber visto un documento y querer encontrarlo otra vez. En otra consulta, el documento estaría ya a la vista y faltaría ubicar la página que debe citarse. La posibilidad de generar una respuesta aparecería más tarde, cuando la búsqueda hubiera reunido pasajes que ayudaran a pensar la pregunta inicial.



## Documentos con historias y condiciones distintas

Mi biblioteca estaba compuesta principalmente por publicaciones que había ido adquiriendo a lo largo de mi carrera profesional. Al pensar en AVANCSO, la diferencia de escala se hizo visible también en la naturaleza de los documentos. Su acervo incluye entrevistas, grabaciones, documentos internos y materiales producidos dentro de relaciones de investigación.

La naturaleza de esos materiales condiciona lo que una aplicación puede hacer con ellos. Las publicaciones suelen llegar al acervo después de haber pasado por una editorial y por algún catálogo. Esto ofrece ciertos puntos de referencia para identificarlas, aunque no resuelve por sí mismo sus condiciones de uso. En los materiales producidos durante una investigación, las condiciones del archivo pueden depender de un acuerdo particular que no aparece en la grabación ni en su transcripción.

El permiso para archivar una entrevista no alcanza necesariamente para dividirla en fragmentos y hacerla aparecer en una conversación con IA. Un pasaje leído por separado puede parecer poco sensible. Al relacionarlo con otros documentos, la aplicación puede volver identificable un lugar o una trayectoria que esa lectura aislada no revelaba.

Las decisiones sobre circulación alcanzan también a lo que queda después del original: una transcripción, por ejemplo, puede conservar el contenido aun cuando el PDF o la grabación ya no estén disponibles. Un índice o una representación vectorial pueden mantener ese contenido disponible para una búsqueda. Los respaldos entran en la misma discusión, ya que eliminar el archivo original no borra automáticamente sus derivados.

Durante un piloto, el centro tendría que discutir qué materiales pueden circular con mayor libertad y cuáles requieren otra clase de tratamiento. La clasificación tendría que poder revisarse cuando cambien las condiciones de acceso de un documento o aparezca información nueva sobre él. La pregunta que me parece más difícil es: ¿qué nuevas relaciones estamos autorizados a producir entre los documentos y las personas que aparecen en ellos?

## Dónde circulan los documentos y quién puede decidir sobre ello

Un documento puede tener una condición de acceso determinada y recorrer infraestructuras muy distintas cuando una aplicación de IA trabaja con él. Para seguir ese recorrido, propongo distinguir las condiciones de acceso de los entornos donde ocurre el procesamiento. Los servicios externos, el procesamiento local y los entornos aislados ofrecen posibilidades diferentes, cuya elección dependerá de lo que pueda hacerse con cada material.

**Las condiciones de acceso** establecen bajo qué circunstancias puede consultarse un documento y qué clase de procesamiento admite. Un permiso para archivar una entrevista puede permitir su conservación sin autorizar que una aplicación la divida en fragmentos o la utilice para generar una respuesta. Esas condiciones pueden cambiar con el tiempo y obligar a revisar un recorrido que antes estaba permitido.

**Los servicios externos** intervienen cuando la aplicación envía una consulta y los fragmentos recuperados hacia un modelo que funciona fuera de la institución. Para utilizar esta ruta con un documento de circulación abierta, todavía habrá que comprobar que sus condiciones de

uso y las reglas del servicio permitan ese procesamiento.

**El procesamiento local** ocurre dentro de una computadora o un servidor administrado por la institución. Los documentos permanecen allí mientras el modelo trabaja con ellos. La institución puede intervenir directamente en ese entorno cuando necesita revisar lo ocurrido o cambiar la manera en que se procesa un material.

**Los entornos aislados** son una forma más estricta de procesamiento local y funcionan sin conexión a internet. Estos entornos pueden ser necesarios cuando la circulación abierta en internet de una copia o de una transcripción presente diversos riesgos.

Con todo, la institución tendrá que decidir qué recorrido corresponde a cada material y quién puede modificarlo cuando cambien sus condiciones de acceso. La capacidad de tomar esas decisiones y sostenerlas con el tiempo es el problema que abordaré a continuación mediante las nociones de soberanía y gobernanza.

## Soberanía como margen de decisión

Utilizo aquí la palabra **soberanía** para hablar de la capacidad de una institución de tomar decisiones sobre los documentos que custodia y de sostenerlas con el tiempo. Esa capacidad se pone a prueba cuando cambia la situación de un documento o cuando una herramienta deja de funcionar y alguien tiene que intervenir.

El lugar donde permanecen los materiales es una primera expresión de ese margen. Una institución puede conservar sus documentos dentro de una infraestructura que administra y decidir qué partes de ellos pueden circular. Si cambian las condiciones de un material, las personas que conocen la historia del acervo pueden intervenir en la forma en que la aplicación lo procesa.

Por **gobernanza** entiendo la manera en que la institución establece quién puede tomar esas decisiones y cómo podrán revisarse después. Un permiso puede cambiar cuando el documento ya ha sido procesado. También puede aparecer información que obligue a reconsiderar su descripción. La aplicación tendría que permitir que alguien modificara esas decisiones y dejar algún registro de lo ocurrido. Sin ese procedimiento, una clasificación puede quedar fijada dentro del sistema aunque ya nadie esté en condiciones de justificarla.

Ya que los modelos de IA y los proveedores de servicios externos cambian, la biblioteca debería conservar sus documentos, sus metadatos y las relaciones entre ellos aun cuando hubiera que reemplazar el modelo que genera las respuestas. Algunos modelos permiten descargar sus pesos, es decir, los parámetros numéricos que aprendieron durante su entrenamiento<sup>8</sup>. Esto puede hacer posible ejecutarlos dentro de la institución. Pero los pesos son apenas una parte de la infraestructura. Todavía hace falta una computadora adecuada y personas que puedan mantener el sistema.

<sup>8</sup> Los pesos son valores numéricos que se generan durante el entrenamiento de un modelo. Son fundamentales porque constituyen la "memoria" o "inteligencia" del sistema de IA. Mientras que algunos modelos son abiertos y permiten descargarlos para usarlos en computadoras locales de forma privada, otros (como ChatGPT o Claude) son cerrados y solo pueden utilizarse a través de sus sitios web o servicios en la nube.

Por cierto, las dependencias no desaparecen cuando la aplicación se instala localmente. La infraestructura seguirá vinculada con quienes fabrican los equipos y con las comunidades que desarrollan el software. La soberanía nombra un margen relativo de autonomía dentro de esas relaciones. Para un centro de documentación, ese margen puede determinar si la aplicación acompaña el cuidado de los materiales o termina imponiéndoles condiciones definidas en otro lugar.

## Los documentos todavía tienen que entrar

Mi experiencia comenzó con una biblioteca formada, en buena medida, por archivos PDF. AVANCSO tendría otro punto de partida. Una parte importante de sus materiales permanece en estantes, cajas, gabinetes y carpetas cuyo orden forma parte de la historia de trabajo del centro.

En las conversaciones sobre AVANCSO, algunas personas señalaron que escanear el acervo podía resultar demasiado costoso. Mi propio experimento cambió cuando compré un escáner vertical. En condiciones adecuadas, el aparato puede capturar las páginas de un libro de trescientas páginas en unos diez minutos. El equipo aceleró esa parte del trabajo y ahora me permite incorporar a la biblioteca literatura impresa que había permanecido en mis estantes. La preparación de cada volumen y la revisión de las imágenes siguen requiriendo tiempo. El rendimiento real en AVANCSO tendría que medirse con una muestra de sus propios materiales.

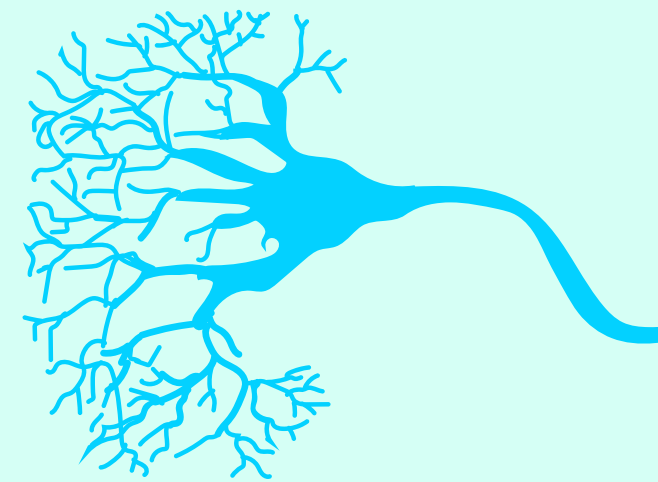
La digitalización podría comenzar con documentos cuyas condiciones de acceso estén claras y que permitan atender alguna de las consultas elegidas para el piloto. Los materiales que requieran otro tratamiento permanecerán fuera de la aplicación mientras el centro discute cómo incorporarlos.

Antes de escanear un documento, quienes preparen el piloto tendrán que situarlo dentro del acervo. Una ficha existente puede ayudar a reconocer su procedencia y las condiciones bajo las cuales se consulta. Si hay varias copias, alguien que conozca la colección tendrá que decidir cuál se digitaliza y cómo quedará vinculada con el original.

El escaneo produce una imagen o un PDF. Para buscar dentro de ese material hace falta extraer o reconstruir su texto. Aquí es donde entra el OCR, o reconocimiento óptico de caracteres. Una página inclinada o un apellido mal reconocido pueden afectar después una consulta. El registro del procesamiento tendría que indicar la calidad del texto y permitir una nueva revisión cuando aparezca un problema.

El OCR entrega el texto con el que trabajarán las aplicaciones. La siguiente tarea consiste en mantener cada página vinculada con el documento del que salió y conservar algún registro de las transformaciones que recibió.

## Una base que permita reconstruir el recorrido



La relación entre cada página y su documento necesitaba una forma de registro. También tenía que saber qué transformaciones había recibido el texto después del OCR. Una consulta ordinaria a un modelo podía devolver de inmediato el nombre de un autor o la fecha de un acontecimiento. En mi biblioteca quería que la misma consulta permitiera volver hasta el documento y la página que sostenían la respuesta.

Comencé por organizar la biblioteca en archivos JSONL, abreviatura de *JSON Lines*<sup>9</sup>. Es un formato de texto en el que cada línea contiene un registro independiente. Una línea podía guardar el texto de una página y la referencia al PDF de origen. Esta organización resultaba suficiente mientras todavía podía reconstruir con relativa facilidad lo que había ocurrido con cada documento.

El crecimiento de la colección volvió más difícil seguir esas relaciones. Cuando una página necesitaba una nueva revisión de OCR, tenía que reconocer qué fragmentos se habían producido a partir de ella para poder actualizarlos. Algo parecido ocurría con la identidad de los documentos. Dos copias del mismo libro podían circular bajo nombres diferentes y requerir tratamientos propios. La consulta de los archivos JSONL comenzó además a hacerse más lenta conforme aumentaba la cantidad de registros.

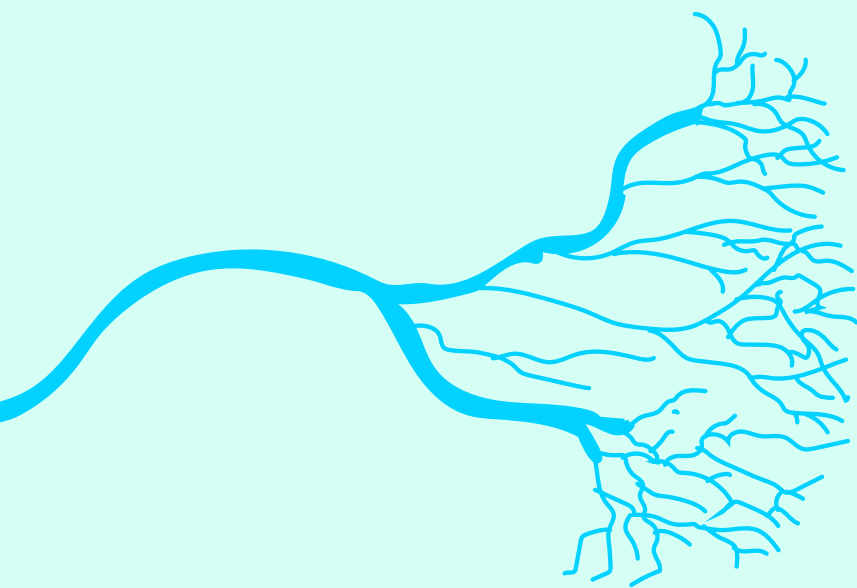
SQLite, una base de datos que puede funcionar dentro de un solo archivo, respondió a esa etapa de mi experimento<sup>10</sup>. Me permitía consultar con mayor rapidez las relaciones entre documentos, páginas y fragmentos sin mantener un servidor aparte. Un centro de documentación puede elegir otro sistema. La solución tendrá que conservar una memoria documental que pueda revisarse cuando cambie un registro o sea necesario reconstruir una parte del procesamiento.

<sup>9</sup> El formato JSONL guarda cada documento o texto en una línea de código independiente. Es el estándar preferido en IA porque permite a los modelos leer y procesar bibliotecas masivas de datos de forma rápida, eficiente y sin saturar la memoria de la computadora.

<sup>10</sup> SQLite es un sistema de bases de datos relacionales que guarda su contenido en un solo archivo de la computadora. Mientras que JSONL es un formato de texto que la IA puede leer de corrido, SQLite permite buscar, filtrar y organizar esa información antes de procesarla.

Los PDF originales permanecían en un espacio separado y la base se refería a ellos mediante un identificador estable calculado a partir de su contenido. Así podía reconocer un archivo aunque cambiara de nombre o de carpeta. Cuando necesitaba reconstruir la base, producía una versión nueva y conservaba la anterior para revisar qué había cambiado.

La trazabilidad es la posibilidad de reconstruir ese recorrido: tome una respuesta cualquiera y trate de volver hasta el PDF y la página. Si el camino se pierde, la respuesta aún no puede presentarse como verificable.



## Formas de buscar dentro de una biblioteca

Cuando las páginas ya habían quedado relacionadas con los documentos de los que provenían, empecé a probar consultas más específicas. La misma biblioteca devolvía resultados distintos según la forma de buscar.

Si recordaba el apellido de una autora, podía escribirlo y encontrar las páginas donde aparecía. Esa era la búsqueda determinística que había intentado construir desde el principio. La aplicación comparaba lo escrito con los metadatos y el texto de las páginas. Yo podía utilizarla también cuando recordaba una frase de manera incompleta y quería volver al lugar donde la había leído.

El problema cambiaba cuando trataba de seguir un concepto entre autores que no utilizaban la misma palabra. Fue ahí donde empecé a usar la búsqueda semántica, que emplea modelos de IA para acercar pasajes relacionados aunque el vocabulario sea distinto. La búsqueda semántica podía mostrar un fragmento que parecía relevante, pero que perdía fuerza cuando lo releía en la página donde aparecía.

Los metadatos entraban en la consulta cuando ya conocía algo del material que buscaba, por ejemplo, el nombre de una autora o un período de publicación permitían reducir el corpus antes de recuperar pasajes. En este documento, corpus nombra únicamente al conjunto de materiales que una aplicación ha sido autorizada a procesar. El acervo puede contener otros documentos que han quedado fuera de ese conjunto.

El resultado de estas consultas seguía siendo un grupo de documentos o pasajes para leer. La posibilidad de pedir una respuesta aparecería después, cuando la búsqueda hubiera reunido materiales que valiera la pena poner en relación.

# Cuando la biblioteca empieza a responder

Las búsquedas que había construido podían llevarme hasta un documento o una página. Algunas consultas me dejaban con varios pasajes cuya relación todavía quería explorar. Fue para trabajar con esos materiales recuperados que empecé a utilizar RAG.

RAG es la abreviatura de *retrieval-augmented generation*, o generación aumentada por recuperación<sup>11</sup>. La aplicación recupera algunos pasajes del corpus y se los entrega a un modelo para que redacte una respuesta a partir de ellos. El corpus ya ha sido delimitado por las condiciones de acceso y por la infraestructura desde la que puede procesarse. Lo que el modelo escribe queda ligado a la manera en que esos fragmentos fueron seleccionados.

En algunas consultas bastaba con encontrar el documento o llegar hasta una página determinada. La respuesta generada aparecía cuando los pasajes recuperados tenían que ponerse en relación para seguir una pregunta. Antes de pedir la respuesta, yo necesitaba poder explicar qué esperaba que el modelo hiciera con ellos.

Mis primeros intentos con RAG me obligaron a desconfiar de las respuestas que sonaban demasiado seguras. Algunas respuestas reunían pasajes pertinentes, pero al volver a las fuentes descubrí que una afirmación iba más lejos de lo que el documento permitía sostener.

Seguir el error hacia atrás me llevaba algunas veces hasta el OCR o hasta la identificación del archivo. Un apellido mal reconocido podía dejar fuera una página pertinente. En otros casos, el modelo recibía un pasaje que había perdido parte de su contexto y redactaba a partir de él una interpretación mal fundada.

Quien revisa una respuesta tiene que leer los pasajes recuperados en las páginas de donde salieron. Después puede preguntarse qué relación guardan con la consulta. Una cita correctamente vinculada

<sup>11</sup> Técnica de Inteligencia Artificial que optimiza las respuestas de un modelo de lenguaje (como GPT o Gemini) al conectarlo con una base de conocimientos externa o privada. En lugar de depender únicamente de sus datos de entrenamiento, el sistema busca fragmentos de información relevantes en documentos específicos (como PDF, manuales o bases de datos) y se los entrega al modelo para que genere una respuesta precisa basada estrictamente en esa fuente real.

obliga todavía a comprobar si la página sostiene la interpretación que la acompaña.

Con el tiempo, una parte de esa revisión comenzó a convertirse en pruebas automáticas. La aplicación puede comprobar que las citas correspondan a los pasajes recuperados. También puede detectar cuándo una modificación hace que una consulta conocida deje de encontrar un material que antes recuperaba. Cuando una prueba detecta uno de esos cambios, la respuesta vuelve a revisión. Allí alguien tendrá que leer las fuentes y decidir si la interpretación puede sostenerse.

## Una biblioteca que recuerda cómo se la consulta

El librero aprende también mientras acompaña las consultas. Con el tiempo recuerda qué documentos ayudaron a responder una pregunta y reconoce algunas búsquedas que terminan repetidamente sin resultados. Al pensar en las aplicaciones institucionales, comencé a preguntarme qué parte de esa experiencia podría conservar la biblioteca.

Una investigadora podría guardar el recorrido que siguió junto con una nota sobre lo que estaba buscando y la utilidad que encontró en cada documento. El recorrido permanecería inicialmente en manos de quien lo creó. Después podría compartirse con un equipo de investigación o con otras personas cuando las condiciones de acceso lo permitieran. La biblioteca comenzaría así a reunir una memoria de las relaciones que sus lectores han encontrado entre los materiales.

Las búsquedas fallidas dejarían otro tipo de registro. Una consulta sin resultados podría señalar una parte del acervo que todavía no ha sido digitalizada. También podría revelar que el OCR perdió un nombre durante el procesamiento. Quienes cuidan el acervo podrían volver a esas consultas para decidir qué trabajo documental atender después.

El uso cotidiano podría señalar además registros que necesitan una revisión. Cuando una consulta revelara un nombre mal leído o una descripción insuficiente, la aplicación prepararía una propuesta acompañada de la página o del resultado que la motivó. Esa propuesta llegaría a una bandeja de revisión. Alguien con responsabilidad sobre el acervo decidiría si la modificación debe incorporarse a la memoria institucional.

## Un asistente que todavía lee

La posibilidad de guardar un recorrido de investigación permitiría imaginar una aplicación que pudiera retomarlo varios días después. Denomino **asistente librero** al conjunto de herramientas que acompañaría ese trabajo durante más tiempo que una consulta aislada.

Una investigadora podría regresar, varios días después, a una pregunta sobre tierra y memoria en un período determinado. El asistente abriría el recorrido que había dejado guardado y mostraría los documentos que ya formaban parte de él. A partir de esos materiales, el asistente podría preparar una comparación entre pasajes que tratan el problema con vocabularios distintos.

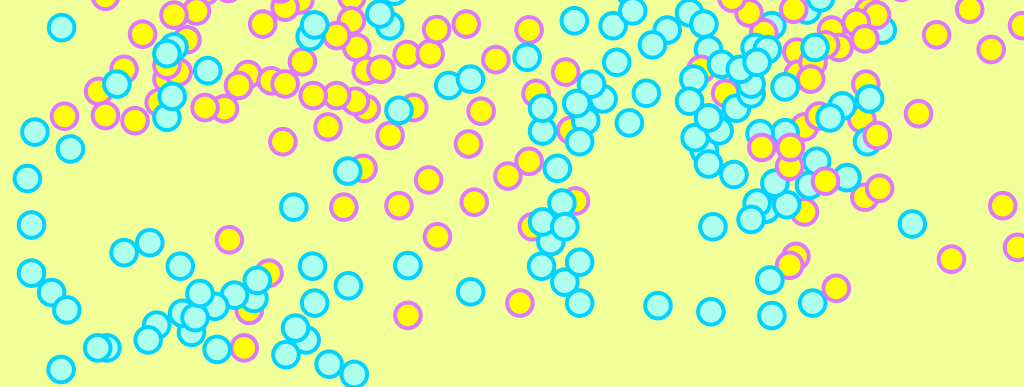
La comparación tendría que conservar el camino seguido por la aplicación. Quien la leyera debería poder abrir las páginas utilizadas y reconocer qué materiales quedaron fuera de la consulta. El asistente mostraría también los límites de acceso que le impidieron trabajar con algún documento.

Una investigación podría conducir hacia un material que todavía no forma parte de la biblioteca. Si fuera de acceso público y su descarga estuviera autorizada, el asistente podría reunirlo en un espacio de ingreso separado del corpus. El documento permanecería allí hasta que alguien revisara su procedencia y decidiera bajo qué condiciones podría utilizarse.

En otros casos, la aplicación podría preparar una solicitud de acceso a un archivo externo. Pedir acceso supone representar a la institución y aceptar condiciones. Una persona autorizada tendría que revisar la solicitud antes de enviarla.

El margen de acción del asistente terminaría en la preparación de esos resultados y propuestas. La memoria documental seguiría bajo el cuidado de quienes tienen responsabilidad sobre el acervo. Si el asistente encontrara una posible corrección, la enviaría a la bandeja de revisión. Una persona decidiría si la propuesta puede incorporarse o si requiere más información.

## Una mesa de trabajo alrededor del acervo



Puedo imaginar el asistente librero como una mesa que alguien reorganiza mientras investiga. Una investigación podría comenzar con una pregunta amplia o con un documento que ya se conoce. Al regresar varios días después, la investigadora podría abrir el recorrido guardado y continuar desde allí.

Cada herramienta atendería una parte de ese trabajo. El orden dependería de la investigación y de los materiales que ya estuvieran disponibles. Los documentos guardados en un módulo podrían abrir una consulta en otro sin obligar a comenzar nuevamente.

### Herramientas de investigación

**Catálogo y recuperación.** El catálogo permitiría encontrar un documento mediante la información conservada en su ficha. Una vez identificado, la recuperación textual podría localizar una palabra o un pasaje dentro de sus páginas. La búsqueda semántica entraría cuando la pregunta necesitara acercar materiales que utilizan vocabularios diferentes.

**Bibliografía comentada.** La aplicación podría preparar una bibliografía a partir de una pregunta y de los materiales recuperados. Cada entrada explicaría la posible utilidad del documento para esa investigación y conservaría el camino hacia el original. La investigadora podría revisar el comentario antes de guardar la bibliografía en su recorrido.

**Conversación con fuentes.** Esta herramienta trabajaría con los documentos que la investigadora hubiera seleccionado. Las respuestas mostrarían las páginas utilizadas y permitirían abrirlas dentro de su contexto. La conversación podría ayudar a comparar pasajes o continuar una pregunta que hubiera quedado abierta durante la lectura.

**Gaveta de investigación.** La gaveta conservaría los archivos PDF que la investigadora quisiera seguir leyendo. Las notas permanecerían junto a los documentos que las motivaron. Al regresar al proyecto, la investigadora encontraría también las búsquedas y bibliografías que hubiera decidido guardar.

### Herramientas para cuidar el acervo

El centro de documentación tendría otra vista de la misma mesa. Desde allí, quienes cuidan el acervo podrían atender los materiales que esperan ingresar y revisar las propuestas producidas durante el uso de la biblioteca.

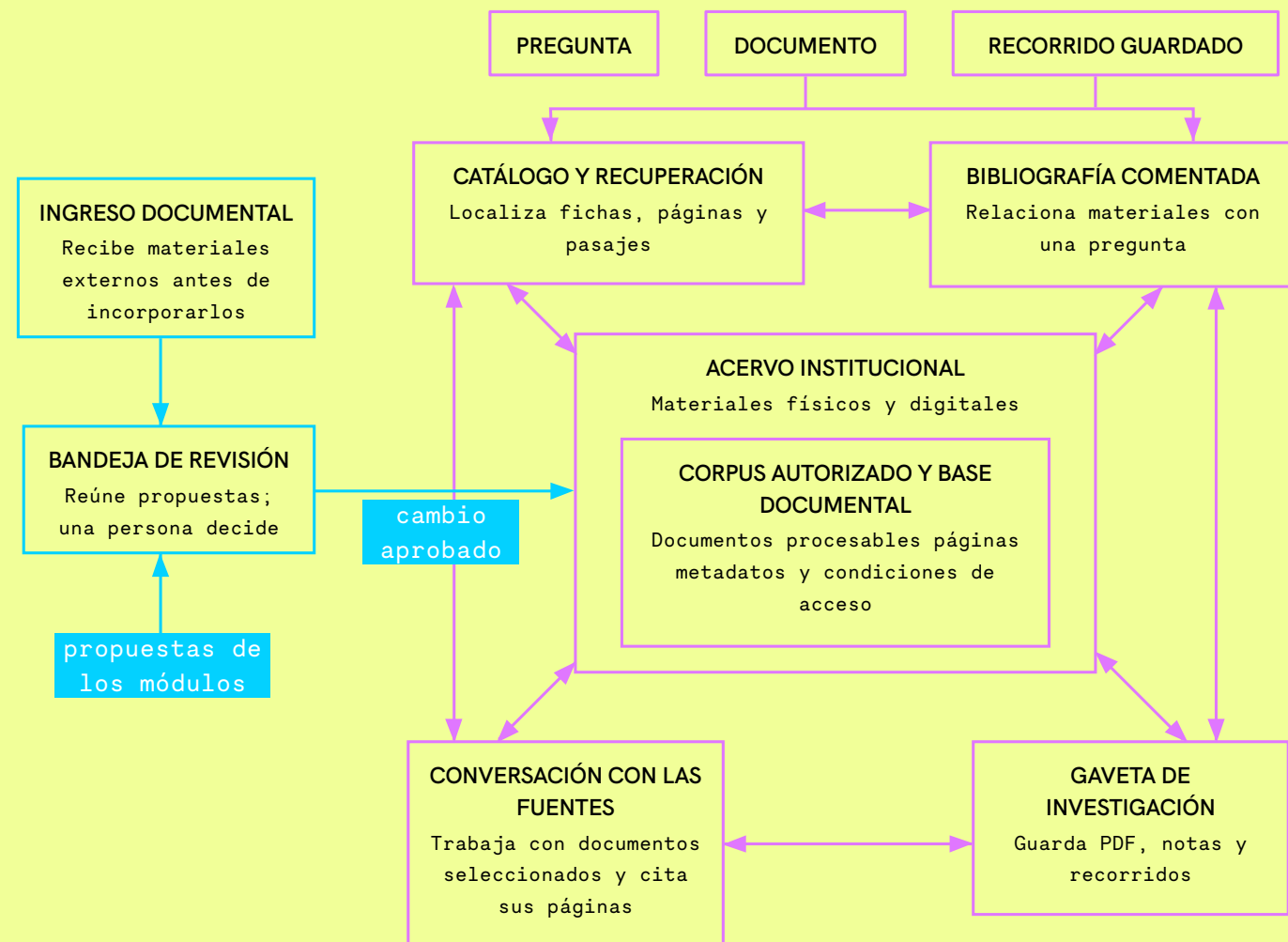
**Ingreso documental.** Un documento localizado fuera de la biblioteca llegaría primero a este espacio. El centro podría revisar su procedencia y establecer bajo qué condiciones puede utilizarse. El material permanecería separado del corpus hasta que una persona autorizara su incorporación.

**Bandeja de revisión.** Las observaciones producidas durante una consulta llegarían a esta bandeja. Una página mal leída por el OCR podría dar lugar a una propuesta de corrección. Si una ficha necesitara más contexto, la propuesta conservaría el documento que la motivó. Una persona con responsabilidad sobre el acervo decidiría qué modificación debe incorporarse y dejaría registrado lo ocurrido.

### Una constelación de herramientas alrededor del acervo

El gráfico (Figura 1) coloca el acervo institucional en el centro de la constelación. Ese acervo incluye los materiales físicos y digitales que permanecen bajo el cuidado del centro. En su interior, el corpus autorizado reúne los documentos que pueden ser procesados por las aplicaciones. La base documental conserva la relación entre esos documentos y sus páginas.

La pregunta, el documento y el recorrido guardado aparecen como posibles puntos de entrada. Las flechas verdes muestran el movimiento de las consultas y sus resultados. Las flechas ocre corresponden a los materiales que esperan ingresar y a las propuestas que requieren una decisión institucional. La flecha que sale de la bandeja de revisión representa un cambio que ya ha sido aprobado.



- rosa: consultas y resultados
- celeste: propuestas y cambios aprobados

Los módulos consultan la misma base. La bandeja de revisión incorpora únicamente los cambios aprobados

**Figura 1.** Los módulos de investigación consultan el mismo corpus autorizado. Las propuestas pasan por la bandeja de revisión antes de modificar la base documental.

Mientras prepara una bibliografía comentada, una investigadora puede advertir que una ficha necesita más contexto. La investigadora enviaría la observación a la bandeja de revisión junto con el documento que la motivó. Si una persona autorizada aprueba la corrección, la ficha revisada quedará disponible para las consultas posteriores.

CONTINUIDAD

## Los README como diario de campo

Como antropólogo, estoy acostumbrado a llevar registro de lo que ocurre durante una investigación. Una nota escrita para aclarar una duda al final de una jornada puede permitir que otra persona vuelva, muchos años después, sobre las condiciones en que se hizo una observación. Al construir la biblioteca descubrí que el trabajo técnico necesitaba algo parecido.

Al principio, una parte de las decisiones había quedado dispersa en conversaciones con modelos y en archivos de código. Cuando regresaba semanas después, tenía que reconstruir por qué había elegido una estructura y qué podía ocurrir si modificaba una parte del sistema. Fue entonces cuando los README comenzaron a adquirir otra función.

Un README es un archivo de texto que suele explicar cómo funciona un proyecto. En mi caso, también empezó a registrar por qué una etapa del trabajo había cambiado. Cuando una solución fallaba, podía dejar anotado lo que había ocurrido y volver después sobre esa decisión. Poco a poco, el README se convirtió en una especie de diario de campo de la infraestructura.

Ese registro permite hacer la prueba de la segunda persona. Quien llegue al proyecto para continuar una tarea tendría que poder reconocer qué se hizo con un documento y dónde quedó explicada esa decisión. Una corrección de OCR o un cambio en la ficha podrían conservarse junto con la razón que llevó a realizarlos.

En un centro de documentación, probaría ese registro entregando a otra persona un documento que ya hubiera pasado por el procesamiento. El README tendría que permitirle reconstruir lo ocurrido y saber dónde intervenir si encuentra un problema. Las decisiones que esa persona no pudiera seguir señalarían los lugares donde todavía hace falta documentar el trabajo.

# Preguntas para poner en marcha un piloto

La primera conversación sobre un piloto puede ocurrir junto a algunos documentos y una consulta que alguien haya hecho recientemente. Esa consulta dará una medida más concreta de lo que la aplicación tendría que hacer y permitirá discutir las siguientes preguntas a partir de materiales conocidos.

**¿Qué consulta queremos acompañar primero?**

Una persona puede recordar un libro sin encontrarlo. En otro momento, puede llegar hasta un documento y necesitar la página donde aparece un pasaje. ¿Qué tendría que poder comprobar antes de utilizar el resultado?

**¿Con qué parte del acervo podemos comenzar?**

La consulta elegida permitirá reconocer qué documentos necesitaría procesar la aplicación. ¿Hay un conjunto cuyas condiciones de acceso ya conozcamos? Antes de digitalizarlo, habrá que situar cada documento dentro del acervo y establecer qué relación conservará con el original (documento físico primario).

**¿Dónde se procesarán los materiales?**

Las condiciones de acceso ayudarán a decidir el recorrido de cada documento. ¿Cuáles pueden utilizar un servicio externo? ¿Qué materiales tendrán que procesarse dentro de la institución o permanecer en un entorno aislado?

**¿Qué trabajo puede sostener el centro?**

Un lote pequeño de documentos permitiría calcular cuánto tiempo requiere prepararlos y revisar el resultado del OCR. Esa experiencia ayudaría a estimar el equipo disponible y el trabajo que las personas responsables podrán mantener después del piloto.

**¿Qué recorrido necesitamos reconstruir?**

Cuando la aplicación encuentra un pasaje, ¿podemos volver al PDF y a la página de donde salió? Si aparece un error, ¿sabremos qué ocurrió durante el procesamiento y qué registros dependen de esa página?

**¿Hasta dónde debe llegar esta primera aplicación?**

Una consulta puede resolverse mediante el catálogo o una búsqueda dentro del texto. Por otro lado, algunas preguntas requerirán reunir pasajes y trabajar con ellos dentro de una conversación con fuentes. La pregunta clave entonces es ¿qué función necesita la consulta que elegimos para el piloto?

**¿Qué funciones pueden esperar?**

El asistente librero podría retomar una investigación durante varios días o preparar el ingreso de un documento externo. El piloto puede conservar esas posibilidades como un horizonte sin construirlas todas en la primera etapa. Relacionada con la pregunta anterior, debemos también contestar ¿qué tarea necesitamos probar ahora y cuál puede desarrollarse después?

**¿Quién decidirá y quién podrá continuar el trabajo?**

El centro tendrá que reconocer quién puede autorizar la incorporación de un documento o aprobar una corrección. Los cambios en la infraestructura quedarán registrados para que otra persona pueda reconstruirlos. La prueba será pedirle que continúe una tarea y observar qué decisiones todavía dependen de quien construyó la primera versión.

# Una infraestructura que sale de mi computadora

Mi experimento personal ya cuenta con un catálogo y con búsquedas textuales y semánticas. La base mantiene vinculados los documentos con sus páginas. También he probado un RAG y algunas verificaciones automáticas para revisar las respuestas que produce.

El paso hacia AVANCSO plantea otra clase de piloto. El centro conserva materiales reunidos durante décadas, con historias de circulación y condiciones de acceso que no coinciden necesariamente con las de mi biblioteca. Las personas que lleguen a consultarlo tampoco compartirán mi recuerdo de los documentos. La aplicación tendrá que aprender a orientarlas sin desprender los materiales de esas relaciones.

Estas páginas pueden funcionar como un heurístico para conversar con quienes enfrentan inquietudes semejantes. Uso esa palabra para nombrar un conjunto provisional de preguntas que ayuda a reconocer decisiones antes de que queden escondidas dentro de una herramienta. Cada centro tendrá que responderlas desde la historia de su acervo y junto con las personas que pueden asumir responsabilidad por él.

La prueba comenzará cuando estas herramientas trabajen con un acervo cuya memoria no es la mía y con las personas que han participado en su cuidado. Entonces será posible observar si una consulta conserva el camino hacia los documentos y si el centro mantiene la capacidad de decidir sobre ellos.

# Una guía de pasos final

Resumen de pasos para preparar aplicaciones de IA en centros de documentación

**PRINCIPIOS PARA COMENZAR**

Define metadatos, accesos y trazabilidad desde el inicio tanto para el original como para sus derivados (transcripciones, índices e IA); así, si las reglas de acceso cambian, podrás rastrear e intervenir de inmediato sobre cualquier copia existente.

Digitalizar relaciones, no solamente documentos. La representación digital debe conservar la relación con el original, su procedencia y las condiciones bajo las cuales se consulta.

Guarda siempre el escaneo original como una copia maestra: sepárala de las versiones de trabajo que se generan a partir de ella. Así, siempre podrás volver a generar nuevos derivados desde la imagen máster original.

Trata el OCR como un resultado que puede contener errores. El texto extraído debe permanecer vinculado con la imagen y la página de origen, y el registro del procesamiento debe permitir una nueva revisión.

Garantiza que cualquier resultado de búsqueda o IA se pueda verificar regresando al documento y a la página exacta de origen; una vez establecida esa trazabilidad, se pueden agregar herramientas como RAG o búsqueda semántica como vías complementarias sin depender de una sola.

**Documenta para la continuidad**

Registrar las decisiones técnicas e hitos del proyecto mediante archivos README concebidos como un “diario de campo”, permitiendo que otras personas puedan dar continuidad a la infraestructura



Por: Dr Alejandro Flores-Aguilar

Gracias al equipo que trabajó en el AI Learning Labs: Andrés Vázquez, Kannika Thaimai, Leticia Coelho, Lucas Pretti, Patricio del Boca, Renata Ávila, Solana Larsen.

Editora: Paz Peña

Ilustraciones y diseño: Constanza Figueroa Bustos

**Octubre de 2026**

Estilo de cita recomendado: Alejandro Flores-Aguilar. (2026). Una biblioteca que ayuda a investigar un acervo – Preguntas y decisiones para preparar aplicaciones de IA en centros de documentación desde la perspectiva de un antropólogo. Open Knowledge Foundation. <https://10.5281/zenodo.23067760>

Open Knowledge Foundation (OKFN) es una organización sin fines de lucro registrada en Inglaterra y Gales con el número de empresa 05133759. [okfn.org](https://okfn.org)

La apertura es nuestra fundación.

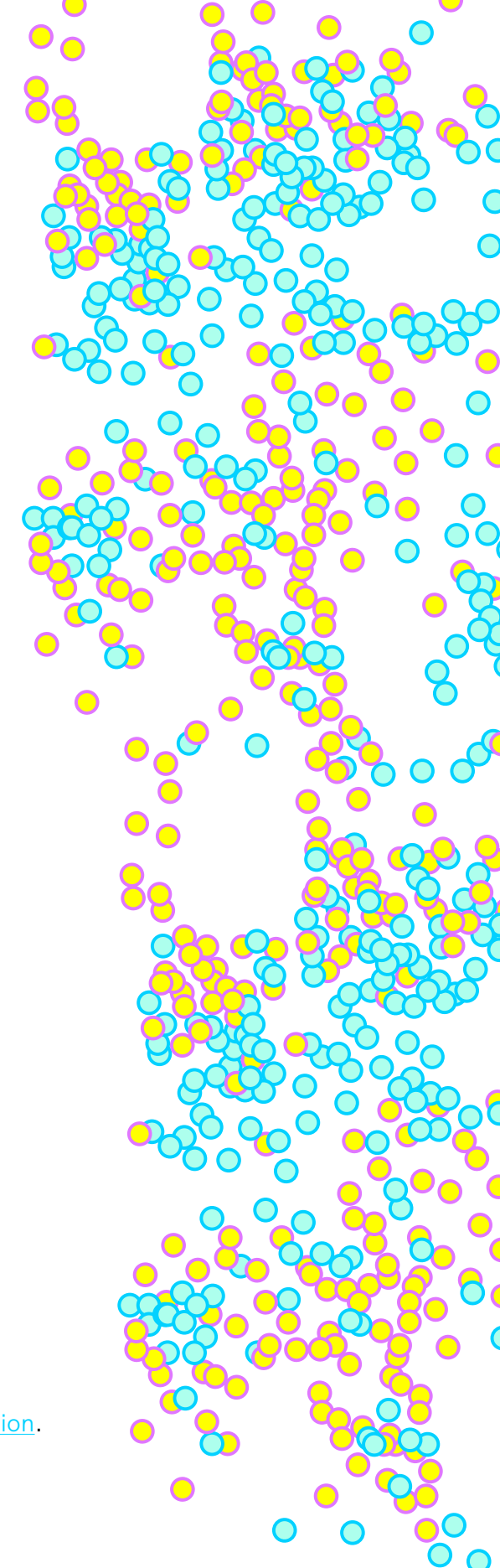


Agradecemos el apoyo de la [Patrick J. McGovern Foundation](https://patrickjmcgovernfoundation.org).

Conozca más sobre sus programas de financiación [aquí](#).



Esta publicación está licenciada bajo una Licencia Creative Commons Atribución 4.0 Internacional ([CC-BY 4.0](https://creativecommons.org/licenses/by/4.0/)).





AI

Learning

Labs

## #AILEarningLabs

Catalizando el aprendizaje y desarrollando métodos replicables para ayudar a las organizaciones a desarrollar habilidades en IA, usar la IA de manera responsable y desarrollar sus propios proyectos de IA.

Más información:  
[okfn.org/ai-learning-labs](https://okfn.org/ai-learning-labs)



### Participe

Ejecuta una prueba piloto:  
¿Interesado en probar esto?  
Contáctanos en [info@okfn.org](mailto:info@okfn.org)

Únete a la conversación: ¿Tienes lecciones o preguntas de tu propio trabajo? Conecta con otros profesionales en nuestro [Open Knowledge Forum](#).



Open Knowledge