

SOME IDEAS DRAWN FROM EXPERIENCE

A library that supports research into a collection

Questions and decisions
for preparing AI
applications in
documentation centres
from an anthropologist's
perspective

Dr Alejandro Flores-Aguilar
AVANCSO



Open Knowledge

GLOSSARY

AVANCSO: The Association for the Advancement of the Social Sciences in Guatemala (AVANCSO) is a social science research centre that began operations in January 1986, following the political opening that led to the establishment of an elected civilian government in Guatemala.

CHATBOT: A computer program designed to simulate a conversation with people via text or voice.

JSONL (JSON Lines): It is a text format in which each line contains a separate record. It can be useful for organising and processing collections of documents straightforwardly and for reading records one by one. Still, it is not an artificial intelligence standard in itself, nor does it replace a document database when you need to consult and maintain more complex relationships between documents, pages, and fragments.

METADATA: A set of data that describes, explains, or provides context for other data. It is commonly called “data about data.” This information can indicate who created a file, when it was modified, what it is about, what format it is saved in, or how large it is. Metadata facilitates the classification and quick retrieval of documents, images, or files, optimising internal search processes within organisations and computer systems.

OPTICAL CHARACTER RECOGNITION

(OCR): The process by which text is extracted or reconstructed from an image of a page. The result is a machine-generated transcription that may contain errors depending on the image quality, typeface, language, page layout, and condition of the document. Therefore, retain the original image, keep it linked to the extracted text, and allow for further revisions if any problems arise.

README: It is an informative document that accompanies a software project, an application or a repository to explain its purpose, content, history and use. It is usually the first file a developer working on a project reads.

Retrieval-Augmented Generation

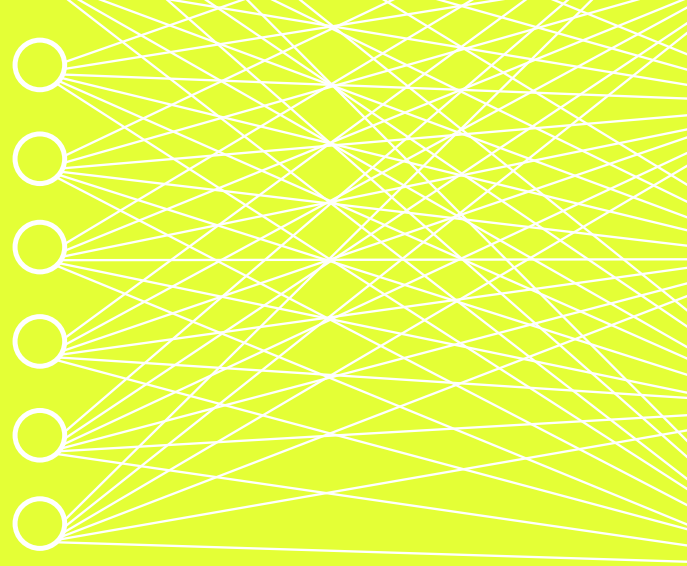
(RAG): This artificial intelligence technique connects a language model to an external or private knowledge base. Rather than relying solely on its training data, the system retrieves relevant snippets from specific documents and feeds them to the model so it can draft a response based on them. Retrieval alone does not guarantee an accurate response or one strictly limited to the sources; that is why it is necessary to retain the path to the documents and check whether they support the interpretation produced.

SANDBOX: An isolated testing environment that lets users run programs or files without affecting the application, system, or platform they run on.

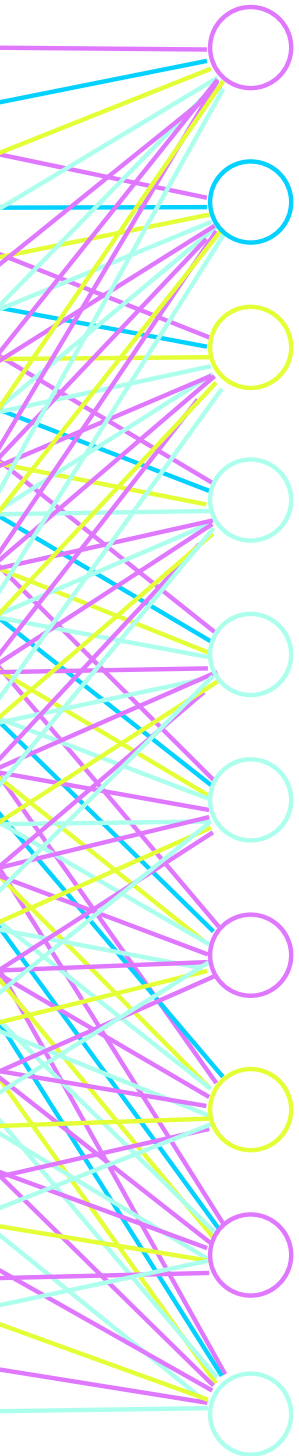
SEMANTIC SEARCH: A search engine technology that interprets the meaning of words and phrases. Semantic search results return content that matches the meaning of a search query, rather than content that literally matches the words in the query.

SQLite: A database management system that runs in a single file on a computer. Unlike a collection of JSONL records, it lets you query, filter, and maintain relationships between structured data without managing a separate server. This may be sufficient for an experiment or certain collections; a documentation centre will need to choose a solution that it can maintain and that preserves the necessary relationships between documents, pages, snippets and transformations.

TEXTUAL MATCH: Also known as exact search or keyword search, this is an information retrieval method in which the system locates character strings that exactly match the text entered by the user.



What these pages are for



These pages are intended for people who work in documentation centres (hereinafter 'centres') and are asking what it takes to prepare a collection for artificial intelligence applications. The text offers a starting point for organising the first decisions involved in a pilot. This question guides it: What would a centre need to do for an application to support research into its materials, while allowing people to return to the documents and pages, on which its results are based?

This guide starts with my experience as an anthropologist building a digital library from my own collection using a local AI model that runs on my computer. Over the years I gathered books, articles, chapters, interviews, and other documents that accumulated on my computer. As the collection grew, the way I had organised the files until then was no longer enough. The problems I encountered were quite simple: sometimes I remembered the title of a text, but the filename bore no obvious relation to it, and the search engine built into my operating system couldn't find it either. I could spend a considerable amount of time looking for a document that I knew was on the computer. Those interruptions gradually slowed down my research.

Since I was already running a local AI model, I thought I could build a small application that read the PDFs and gave them more useful names. It worked badly. To suggest a name, the program first had to recognise the document in front of it. The metadata¹ might be missing or incorrect. When the PDF came from a scan, the program sometimes could not even read the title.

That first experiment changed the question I was asking. The program first had to recognise which document each file contained. If its pages existed only as images, it also had to convert them into text. Afterwards, each page had to retain its relationship to the original document. Perhaps the same infrastructure would allow me to find specific passages in articles I had read many years earlier or follow the trajectory of a concept through a broader body of literature.

¹ Metadata describes data's content, context, and provenance in a standardised, structured way. See the 'Glossary' section.

This is how I began to build what might be called a bibliographic chatbot.² The phrase was appealing, although it soon proved misleading. The conversation window drew all the attention. Much of the work, however, took place before anyone typed a question.

I gradually developed two search methods. One was deterministic and worked with titles, authors, dates and exact matches. The other tried to retrieve relationships between materials even when they did not share the same words. For months, the library was mainly a set of folders, extracted pages, and tables attempting to remember which document each item had come from.

After speaking with the Chief Executive Officer of the Open Knowledge Foundation, I began to consider taking these early experiments to the Documentation Centre of the Association for the Advancement of the Social Sciences in Guatemala, AVANCSO. My collection could serve as a sandbox,³ a space that tolerated my provisional decisions and my poorly named folders while I learnt.

But AVANCSO shifted the question: a documentation centre required a different kind of attention. AVANCSO's collection has another kind of memory. Its materials were gathered over decades, and different people have helped look after them. Some of the knowledge about relationships between the documents remains embedded in the practices of those who have worked with them. An institutional application would have to learn to work with that distributed memory.

The exchange with Open Knowledge Foundation suggested that some of these lessons might be useful to other organisations. A tool built from my library offers a starting point. Its categories would have to be discussed when it encountered another collection. Each centre will need to decide which problem it wants to address; that decision will make it possible to establish which relationships it must preserve and which application it might develop under its own conditions. In this way, the question began to be shared.

² Computer software that simulates conversations with people via text or voice.

³ An isolated test environment that lets users run programmes or execute files without affecting the application, system, or platform they run on.



Reading routes

The sections follow the path from my personal library towards the possibility of an institutional pilot. The contents page traces that path in full. A centre can also begin with the problem it needs to discuss. The following diagram may help:

Index in order

If much of the collection remains on paper, these sections examine the decisions that precede processing.

What these pages are for	4
From a personal library to a shared question	7
Begin with an enquiry	9
Documents with different histories and conditions	10
Where documents circulate and who can decide about them	11
Sovereignty as criteria for decision-making	12
The documents still need to be processed	14
A database that can reconstruct the path	15
Ways of searching within a library	17
When the library begins to answer	18
A library that remembers how it is consulted	20
An assistant that still reads	21
A worktable around the collection	22
README files as a field journal	25
Questions for launching a pilot	26
An infrastructure that leaves my computer	28
A step-by-step guide to the final stages	29

If the centre has digitised materials, these sections address traceability, retrieval, and generated responses.

If the discussion concerns the applications that might be built, these sections develop that horizon.

If the centre wants to define a pilot and sustain it over time, these sections bring together the institutional decisions that will need to be made.

From a personal library to a shared question



The experiment I ran to consult my library, lets me find documents and search within their pages. It can also retrieve passages on the basis of semantic proximity and generate responses from them. The database preserves the path back to the sources, and several automated checks help me review what happens when a response loses that connection.

How did I build it? I outline that process on these pages to provide an overview.

My starting point—and that of a centre—is a recent enquiry that needs addressing. For example, someone may remember a book but be unable to find it. They may already have the document in front of them but need the page on which a passage appears. Working from a recognisable difficulty helps clarify what function the application should perform and what the person would need to verify before using what they found. The chosen enquiry helps delimit the part of the collection that would need to be processed.

The centre would have to review the access conditions attached to those documents and decide which infrastructure they may pass through. If the materials are on paper, scanning will produce images that Optical Character Recognition (OCR)⁴ must convert into text. The documentary base will have to preserve the relationship between that text, the page, and the original file.

Once those relationships have been recorded, the form of search can respond to the kind of question being asked. A textual match⁵ is useful for locating a surname or a phrase. Semantic search⁶ can bring

4 OCR (Optical Character Recognition) extracts and reuses data from scanned documents, camera images, and image-only PDFs. See the 'Glossary' section.

5 A method in which the system searches for the exact words you type in a document or on the internet. See the 'Glossary' section.

6 An advanced method that looks for the meaning and intention behind your words, rather than simply matching exact letters. See the 'Glossary' section.

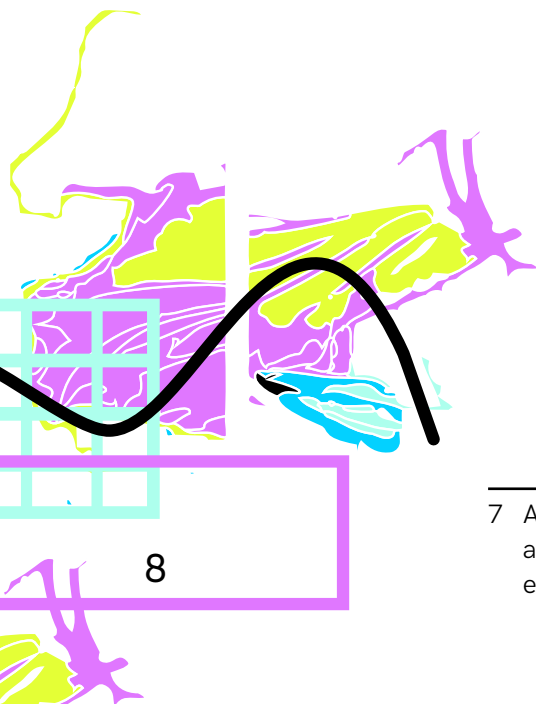
together passages that address a problem using different vocabularies. For some enquiries, the retrieved passages can be given to a model so that it can draft a response from them. This operation is called Retrieval-Augmented Generation (RAG)⁷. At a later stage, the application might retain a research path and resume it several days later.

That is the current extent of the infrastructure I have built.

As we shall see later in this document, aspects such as the 'librarian-like assistant' form part of the projective phase of the AVANCSO pilot: the worktable and the constellation of tools provide a way of imagining how it might accompany a longer research process. The institutional infrastructure for AVANCSO will also need to be designed and tested.

As I mentioned earlier, AVANCSO changes the scale of the question compared to my personal library. Its documentation centre holds materials gathered over several decades, and the people who consult it will arrive with knowledge different from my own. Access conditions will have to be decided with those who know the history of the collection and have authority over its documents. The pilot must also remain in the hands of people who can continue it when the tools change or when the person who built the first version leaves.

I use heuristics to refer to the provisional set of questions gathered in these pages. It can support conversations with other centres considering similar work and help identify decisions that often remain hidden inside an application. The discussion would need to take place alongside specific documents and the people responsible for caring for them. Each project will have to calculate the cost of preparing its materials and establish what infrastructure it can maintain. Those answers will allow it to define a pilot suited to its circumstances.



⁷ An artificial intelligence technique that optimises the responses of a language model (such as GPT or Gemini) by connecting it to an external or private knowledge base. See the 'Glossary' section.

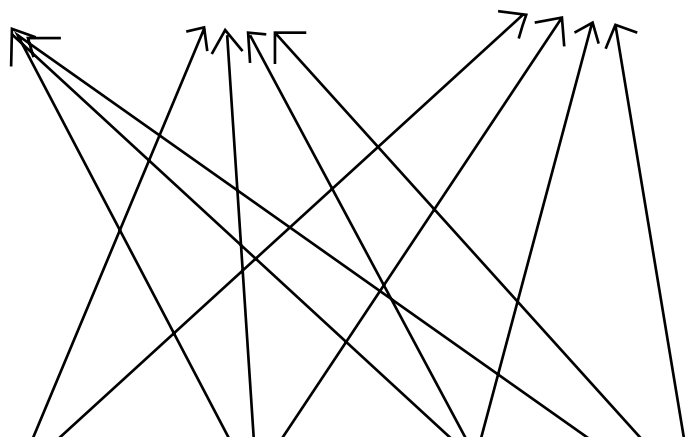
Begin with an enquiry

My library had an advantage that would be difficult to transfer to a documentation centre. I could recognise a mistyped title or notice that a response did not correspond to what I remembered of a work. That personal memory compensated for some of the application's weaknesses. People consulting AVANCSO's collection would arrive with their own references and would need guidance even when they did not know in advance which document to look for.

This difference led me to think of a good librarian. Someone who has known a library for a long time can listen to a question that is still imprecise and guide a person towards documents they may not have considered. The librarian also knows when it is better to return to the catalogue, open a box, or look for a page before venturing an answer.

The figure of the librarian suggested a way to begin designing the application. Before choosing an AI model, I would observe several searches as they happened. I would want to know what the person is trying to do and where they get stuck. A specific difficulty makes it possible to begin with something the centre already recognises and can describe.

Following those enquiries would reveal different needs. Someone might remember seeing a document and want to find it again. In another enquiry, the document might already be in front of them, but the page to be cited would still need to be located. The possibility of generating a response would arise later, once the search had brought together passages that helped develop the initial question.



Documents with different histories and conditions



My library consisted mainly of publications that I had acquired over the course of my professional career. When I began thinking about AVANCSO, the difference in scale also became visible in the nature of the documents. Its collection includes interviews, recordings, internal documents, and materials produced within research relationships.

The nature of those materials conditions what an application can do with them. Publications usually enter the collection after passing through a publisher and a catalogue. This provides some reference points for identifying them, although it does not in itself settle their conditions of use. For materials produced during research, the terms under which they may be archived can depend on a particular agreement that appears neither in the recording nor in its transcription.

Permission to archive an interview does not necessarily extend to dividing it into fragments and making it appear in a conversation with AI. A passage read on its own may seem relatively innocuous. When related to other documents, the application may make a place or trajectory identifiable in a way that the isolated reading did not.

Decisions about circulation also extend to what remains after the original: a transcription, for example, may preserve the content even when the PDF or recording is no longer available. An index or a vector representation may keep that content retrievable. Backups belong to the same discussion, since deleting the original file does not automatically erase its derivatives.

During a pilot, the centre would have to discuss which materials may circulate more freely and which require another kind of treatment. The classification would need to remain open to revision when a document's access condition changes or new information about it emerges. The most difficult question, in my view, is this: What new relationships are we authorised to produce between the documents and the people who appear in them?



Where documents circulate and who can decide about them

A document may have a particular access status yet pass through very different infrastructures when an application works with it. To follow that path, I propose distinguishing access conditions from the environments in which processing takes place. External services, local processing, and isolated environments offer different possibilities; the choice between them will depend on what may be done with each material.

Access conditions establish the circumstances under which a document may be consulted and the kinds of processing it permits. Permission to archive an interview may allow it to be preserved without authorising an application to divide it into fragments or use it to generate a response. Those conditions can change over time and make it necessary to review a path that was previously allowed.

External services are involved when the application sends an enquiry and the retrieved fragments to a model operating outside the institution. Before using this route for a publicly accessible document, it is still necessary to establish that both its

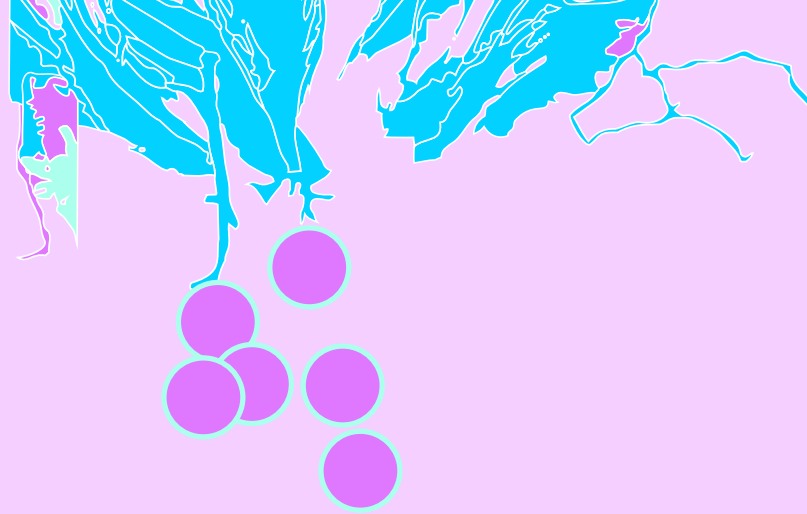
conditions of use and the service's rules permit that processing.

Local processing takes place on a computer or server administered by the institution. The documents remain there while the model works with them. The institution can intervene directly in that environment when it needs to review what happened or change the way a material is processed.

Isolated environments are a stricter form of local processing and operate without an internet connection. They may be necessary when the open online circulation of a copy or transcription poses risks.

The institution will have to decide which path is appropriate for each material and who may change it when its access conditions shift. The capacity to make those decisions and sustain them over time is the problem I address next through the ideas of sovereignty and governance.

Sovereignty as criteria for decision-making



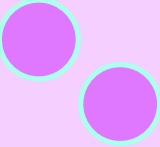
I use **sovereignty** here to mean an institution's capacity to make decisions about the documents in its care and to uphold those decisions over time. That capacity is tested when a document's situation changes or when a tool stops working and someone has to intervene.

The place where the materials are kept is one expression of that room for decision-making. An institution can keep its documents within the infrastructure that it administers and decide which parts of them may circulate. If a material's conditions change, the people who know the collection's history can intervene in the way the application processes it.

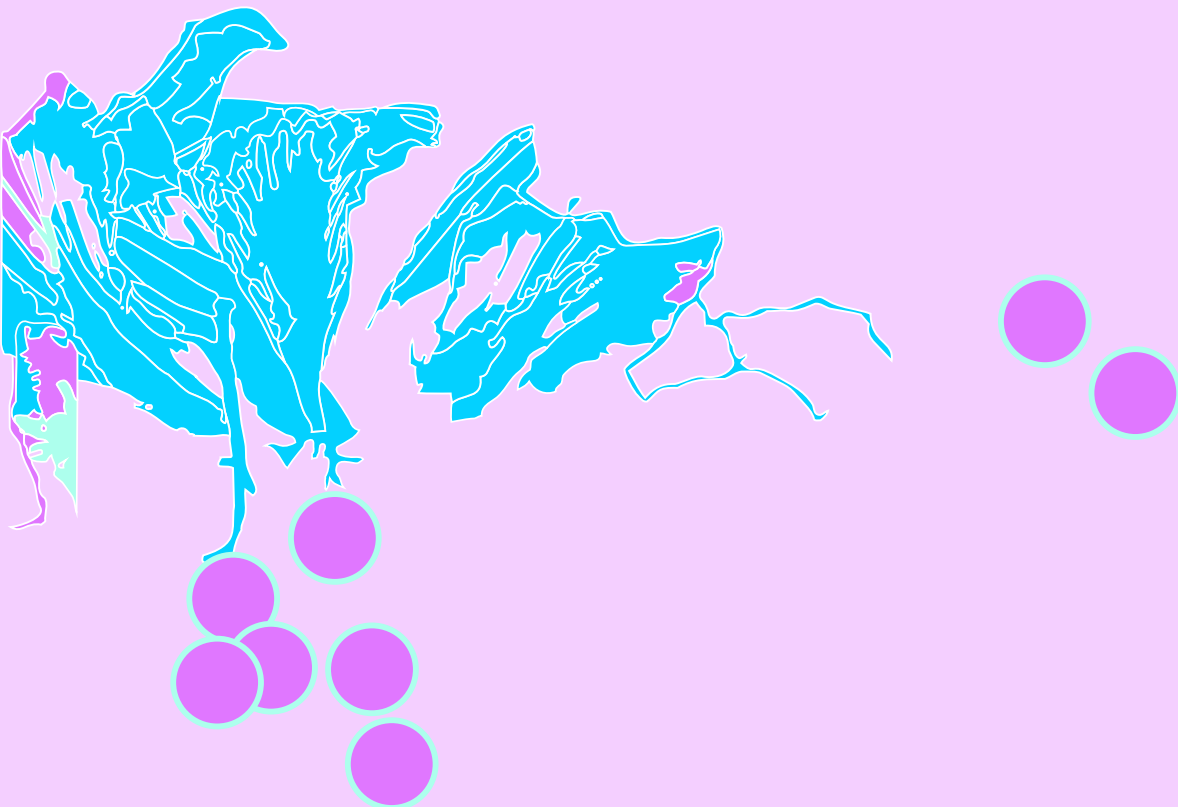
By **governance** I mean the way the institution establishes who may make those decisions and how they can later be reviewed. A permission may change after the document has already been processed. Information may also emerge that requires its description to be reconsidered. The application would need to allow someone to alter those decisions and to leave a record of what occurred. Without such a procedure, a classification may become fixed within the system even though no one is any longer able to justify it.

Because AI models and external service providers change, the library should retain its documents, metadata, and the relationships between them even if the model generating the responses has to be replaced. Some models allow their weights to be downloaded, meaning the numerical parameters learnt during training.⁸ This may make it possible to run them within the institution. The weights, however, are only one part of the infrastructure. A suitable computer and people who can maintain the system are still needed.

⁸ Weights are the numerical values that the model calculates during training. They are fundamental because they constitute the AI system's 'memory' or 'intelligence'. Whilst some models are open-source and can be downloaded for private use on local computers, others (such as ChatGPT or Claude) are closed or proprietary and can only be used via their websites or cloud services.



Dependencies do not disappear when an application is installed locally. The infrastructure will remain connected to the companies that manufacture the equipment and the communities that develop the software. Sovereignty names a relative degree of autonomy within those relationships. For a documentation centre, that room for action may determine whether the application supports the care of its materials or ends up imposing conditions defined elsewhere.



The documents still need to be processed

My experience began with a library made up largely of PDF files. AVANCSO would have a different starting point. A substantial part of its materials remains on shelves and in boxes, cabinets, and folders whose order forms part of the centre's working history.

In conversations about AVANCSO, several people noted that scanning the collection might be too expensive. My own experiment changed when I bought an overhead scanner. Under suitable conditions, the device can capture the pages of a three-hundred-page book in about ten minutes. It accelerated that part of the work and now allows me to add to the library printed literature that had remained on my shelves. Preparing each volume and reviewing the images still takes time. Actual performance at AVANCSO would have to be measured using a sample of its own materials.

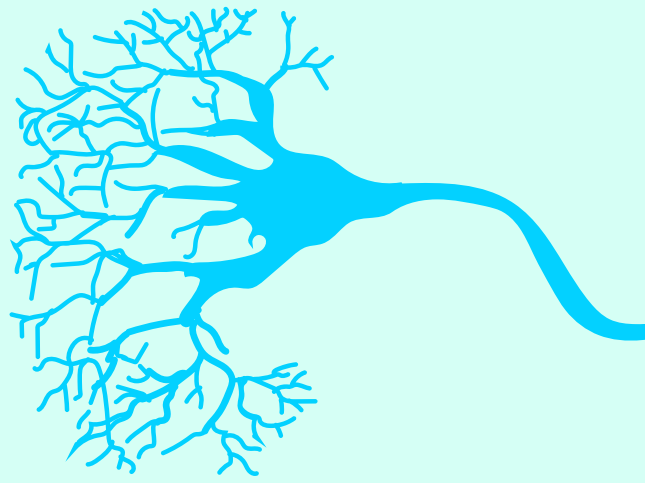
Digitisation could begin with documents whose access conditions are clear and that can address one of the enquiries selected for the pilot. Materials requiring different treatment would remain outside the application while the centre discusses how to include them.

Before scanning a document, those preparing the pilot will have to situate it within the collection. An existing catalogue record may help establish its provenance and the conditions under which it may be consulted. If there are several copies, someone who knows the collection will have to decide which one is digitised and how it will remain linked to the original.

Scanning produces an image or a PDF. Searching within that material requires its text to be extracted or reconstructed. This is where optical character recognition (OCR) enters. A skewed page or a misread surname can later affect an enquiry. The processing record would need to indicate the quality of the text and allow a page to be reviewed again when a problem emerges.

OCR provides the text with which the applications will work. The next task is to keep each page linked to the document from which it came and preserve a record of the transformations it received.

A database that can reconstruct the path



The relationship between each page and its document needed a form of record. I also had to know what transformations the text had undergone after OCR. An ordinary query to a model could immediately return the name of an author or the date of an event. In my library, I wanted the same query to allow me to return to the document and page on which the answer was based.

I began by organising the library in JSONL files, short for *JSON Lines*.⁹ It is a text format in which each line contains an independent record. A line could store the text of a page and a reference to the source PDF. This organisation was sufficient while I could still reconstruct with relative ease what had happened to each document.

As the collection grew, those relationships became harder to follow. When a page needed a new OCR review, I had to identify which fragments had been produced from it so that I could update them. Something similar happened with the identity of documents. Two copies of the same book could circulate under different names and require their own treatment. Querying the JSONL files also became slower as the number of records increased.

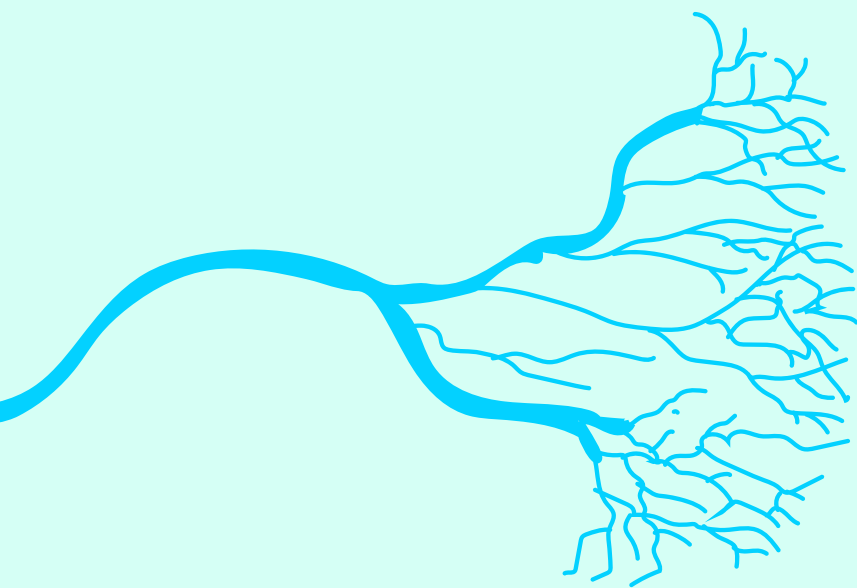
SQLite,¹⁰ a database that can operate within a single file, met the needs of that stage of my experiment. It allowed me to query relationships between documents, pages, and fragments more quickly without maintaining a separate server. A documentation centre may choose a different system. The solution will need to preserve a document history that can be reviewed when a record changes or part of the processing has to be reconstructed.

⁹ The JSONL format stores each document or text on its own line. It is the preferred standard in AI because it lets models read and process massive datasets quickly and efficiently without overloading the computer's memory.

¹⁰ SQLite is a system that stores databases in a single file on the computer. While JSONL is a text format designed to let AI read documents straight through, SQLite lets you search, filter, and organise that information before processing it.

I kept the original PDFs in a separate space, and the database referred to them by means of a stable identifier calculated from their contents. This allowed me to recognise a file even when its name or folder changed. When I needed to rebuild the database, I produced a new version and retained the previous one so that I could review what had changed.

Traceability is the possibility of reconstructing that path. Take any response and try to return to the PDF and the page. If the path is lost, the response cannot yet be presented as verifiable.



Ways of searching within a library

Once the pages had been linked to the documents from which they came, I began testing more specific enquiries. The same library returned different results depending on how I searched.

If I remembered an author's surname, I could type it and find the pages on which it appeared. This was the deterministic search I had tried to build from the beginning. The application compared what I typed with the metadata and the text of the pages. I could also use it when I remembered a phrase incompletely and wanted to return to the place where I had read it.

The problem changed when I tried to follow a concept across authors who did not use the same word. This was when I began using semantic search, which employs AI models to bring together related passages even when their vocabulary differs. Semantic search could retrieve a fragment that sounded relevant but became less persuasive when I reread it on the page where it appeared.

Metadata entered the enquiry when I already knew something about the material I sought; for example, an author or publication period could narrow the corpus before passages were retrieved. In this document, corpus refers only to the set of materials an application is authorised to process. The collection may contain other documents that remain outside that set.

The result of these enquiries was still a group of documents or passages to be read. The possibility of requesting a response would arise later, once the search had gathered materials that were worth bringing into relation.

When the library begins to answer

The searches I had built could take me to a document or a page. Some enquiries left me with several passages whose relationship I still wanted to explore. I began using RAG in order to work with those retrieved materials.

RAG stands for *retrieval-augmented generation*.¹¹ The application retrieves passages from the corpus and gives them to a model so that it can draft a response from them. The corpus has already been delimited by access conditions and by the infrastructure from which it may be processed. What the model writes is tied to the way those fragments were selected.

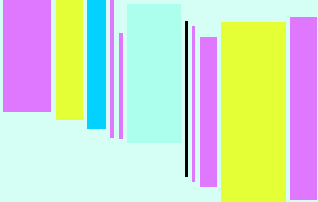
For some enquiries, finding the document or reaching a particular page was enough. I turned to a generated response when the retrieved passages had to be brought into relation in order to pursue a question. Before requesting one, I needed to be able to explain what I expected the model to do with them.

My first experiments with RAG taught me to distrust responses that sounded too certain. Some brought together relevant passages, but when I returned to the sources I discovered that a claim went beyond what the document could support.

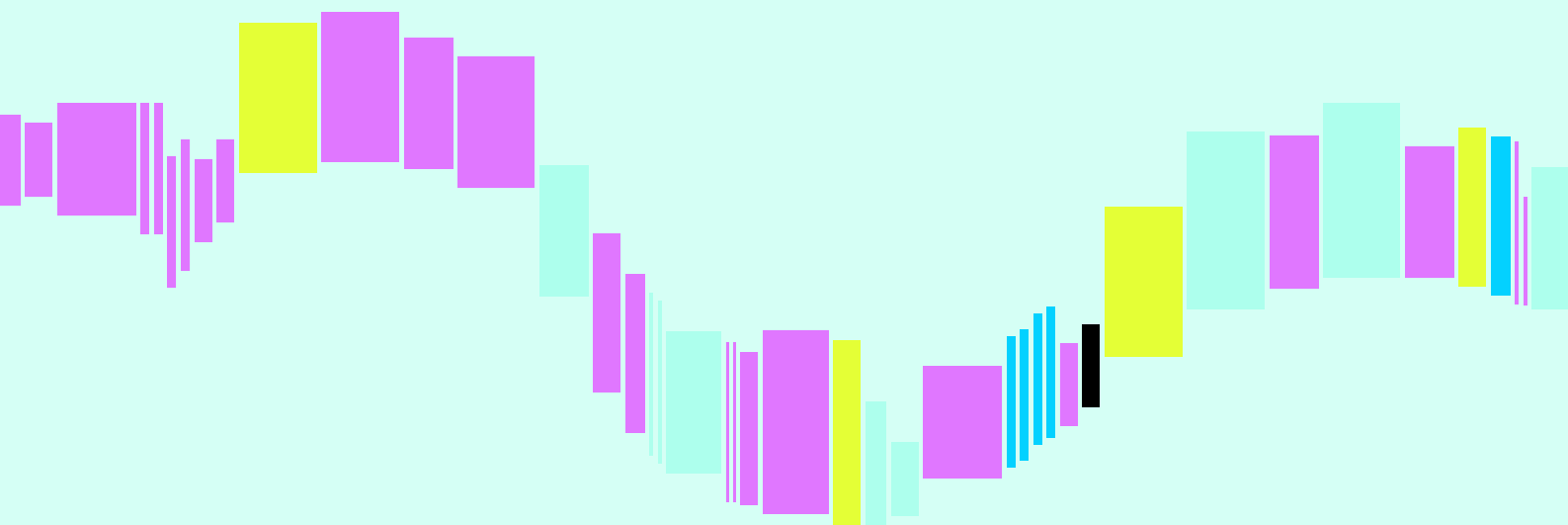
Tracing the error backwards sometimes led me to the OCR or to the identification of the file. A misread surname could exclude a relevant page. In other cases, the model received a passage that had lost part of its context and drafted a poorly grounded interpretation from it.

Anyone reviewing a response has to read the retrieved passages on the pages from which they came. They can then ask how those passages relate to the enquiry. Even a correctly linked citation still requires the reviewer to establish whether the page supports the interpretation that accompanies it.

¹¹ An artificial intelligence technique that optimises the responses of a language model (such as GPT or Gemini) by connecting it to an external or private knowledge base. Rather than relying solely on its training data, the system searches for relevant snippets in specific documents (such as PDFs, manuals, or databases) and feeds them to the model so it can generate an accurate response based strictly on that real-world source.



Over time, part of that review began to take the form of automated tests. The application can check that citations correspond to the retrieved passages. It can also detect when a modification causes a familiar enquiry to stop finding material it previously retrieved. When a test detects one of these changes, the response returns for review. Someone will then have to read the sources and decide whether the interpretation can be sustained.



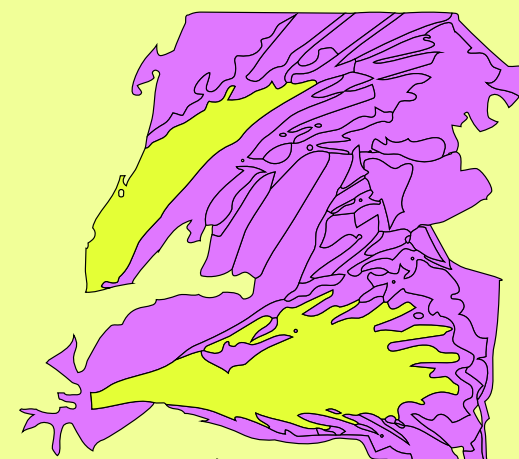
A library that remembers how it is consulted

A librarian also learns while helping with enquiries. Over time, the librarian remembers which documents helped answer a question and recognises searches that repeatedly end without results. When I began thinking about institutional applications, I wondered how much of that experience the library might retain.

A researcher could save the path she followed together with a note about what she was looking for and how each document had proved useful. At first, the path would remain in the hands of the person who created it. It could later be shared with a research team or with other people when access conditions allowed. In this way, the library would begin to gather a memory of the relationships its readers had found between the materials.

Failed searches would leave another kind of record. An enquiry with no results could point to a part of the collection that had not yet been digitised. It might also reveal that OCR had missed a name during processing. Those caring for the collection could return to these enquiries when deciding which documentary work to address next.

Everyday use could also identify records that need review. When an enquiry revealed a misread name or an inadequate description, the application would prepare a proposal accompanied by the page or result that prompted it. The proposal would reach a review tray. Someone with responsibility for the collection would decide whether the change should be incorporated into the institutional memory.



An assistant that still reads

Once a research path could be saved, I could imagine an application returning to it several days later. I use the term **librarian-like assistant** for the set of tools that would accompany this work for longer than a single enquiry.

A researcher might return several days later to a question about land and memory during a particular period. The assistant would open the path she had saved and show the documents that were already part of it. Working from those materials, it could prepare a comparison between passages that addressed the problem using different vocabularies.

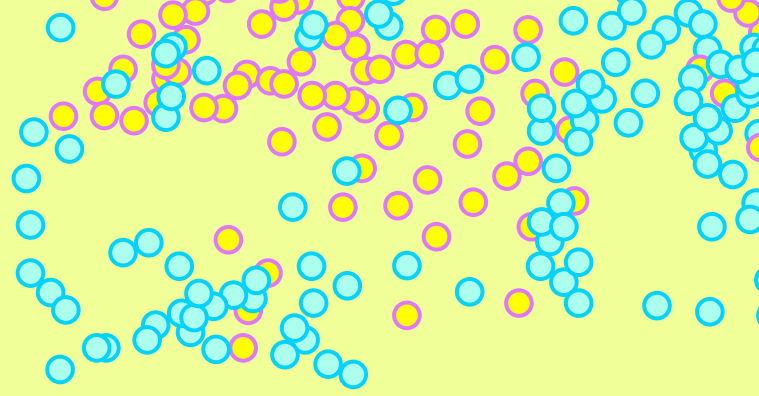
The comparison would have to preserve the path followed by the application. Its reader should be able to open the pages used and recognise which materials had remained outside the enquiry. The assistant would also show the access restrictions that had prevented it from working with a particular document.

A research process might lead to material that was not yet part of the library. If it were publicly accessible and its download were authorised, the assistant could gather it in an intake space separate from the corpus. The document would remain there until someone reviewed its provenance and decided under what conditions it could be used.

In other cases, the application could prepare a request for access to an external archive. Requesting access involves representing the institution and accepting conditions. An authorised person would have to review the request before it was sent.

The assistant's scope for action would end with the preparation of those results and proposals. The documentary record would remain in the care of those responsible for the collection. If the assistant found a possible correction, it would send it to the review tray. A person would decide whether the proposal could be incorporated or whether it required further information.

A worktable around the collection



I imagine the librarian-like assistant as a table that someone rearranges while conducting research. A research process could begin with a broad question or with a document that is already known. When returning several days later, the researcher could open the saved path and continue from there.

Each tool would address one part of that work. The order would depend on the research and the materials already available.

Documents saved in one module could open an enquiry in another without requiring the researcher to begin again.

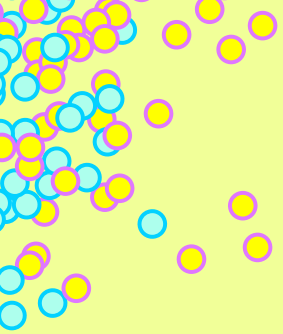
Research tools

Catalogue and retrieval. The catalogue would make it possible to find a document using the information held in its record. Once the document had been identified, textual retrieval could locate a word or passage within its pages. A semantic search could bring together materials that use different vocabularies.

Annotated bibliography. The application could prepare a bibliography from a question and the retrieved materials. Each entry would explain the document's possible usefulness for that research and preserve the path back to the original. The researcher could review the annotation before saving the bibliography to their research path.

Conversation with sources. This tool would work with the documents the researcher had selected. The responses would show the pages used and allow them to be opened in context. The conversation could help compare passages or continue a question that had remained open during reading.

Research drawer. The drawer would hold the PDF files that the researcher wanted to continue reading. Notes would remain alongside the documents that prompted them. On returning to the project, the researcher would also find the searches and bibliographies they had chosen to save.



Tools for caring for the collection

The documentation centre would have another view of the same table. From there, those caring for the collection could attend to materials awaiting entry and review proposals produced through the library's use.

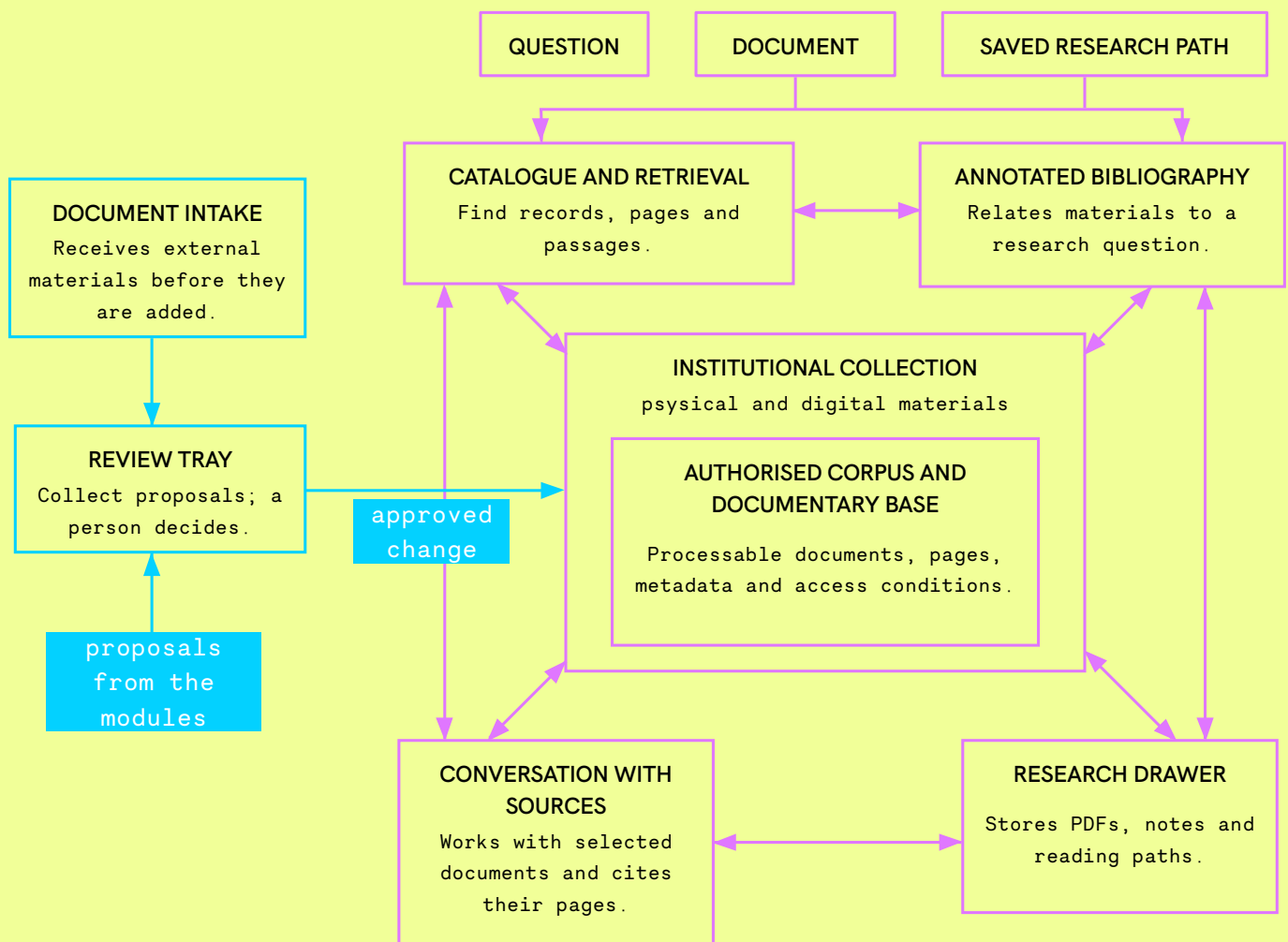
Document intake. A document located outside the library would first arrive in this space. The centre could review its provenance and establish the conditions under which it may be used. The material would remain separate from the corpus until an authorised person approved its incorporation.

Review tray. Observations produced during an enquiry would arrive in this tray. A page misread by OCR could give rise to a proposed correction. If a catalogue record needed more context, the proposal would retain the document that prompted it. A person with responsibility for the collection would decide which change should be incorporated and would leave a record of what occurred.

A constellation of tools around the collection

The diagram (Figure 1) places the institutional collection at the centre of the constellation. This collection includes the physical and digital materials that remain in the centre's care. Within it, the authorised corpus brings together the documents that may be processed by the applications. The documentary base preserves the relationship between those documents and their pages.

The question, the document, and the saved research path appear as possible points of entry. The green arrows show the movement of enquiries and their results. The ochre arrows correspond to materials awaiting entry and proposals requiring an institutional decision. The arrow leaving the review tray represents a change that has already been approved



- pink: queries and results
- celeste: proposals and approved changes

The modules consult the same base. The review tray incorporates only approved changes.

Figure 1. The research modules consult the same authorised corpus. Proposals pass through the review tray before they modify the documentary base.

While preparing an annotated bibliography, a researcher may notice that a catalogue record needs more context. They would send the observation to the review tray together with the document that prompted it. If an authorised person approves the correction, the revised record will be available to later enquiries.

README files as a field journal

As an anthropologist, I am accustomed to keeping a record of what happens during research. A note written to clarify a question at the end of a day may allow someone else, many years later, to return to the conditions under which an observation was made. While building the library, I discovered that technical work needed something similar.

At first, some of the decisions had been scattered across conversations with models and code files. When I returned weeks later, I had to reconstruct why I had chosen a particular structure and what might happen if I altered part of the system. This was when the README files began to take on another function.

A README is a text file that usually explains how a project works. In my case, it also began to record why a stage of the work had changed. When a solution failed, I could note what had happened and return to that decision later. Gradually, the README became a kind of field journal for the infrastructure.

That record makes it possible to carry out the second-person test. Someone arriving at the project to continue a task should be able to determine what had been done with a document and where that decision was explained. An OCR correction or a change to a catalogue record could be preserved alongside the reason for making it.

In a documentation centre, I would test that record by giving another person a document that had already gone through processing. The README should allow them to reconstruct what occurred and know where to intervene if they found a problem. Any decisions they could not follow would indicate where the work still needs to be documented.

Questions for launching a pilot

The first conversation about a pilot can take place alongside several documents and an enquiry that someone has made recently. That enquiry will provide a more concrete measure of what the application would need to do and make it possible to discuss the following questions using familiar materials.

Which enquiry do we want to support first?

Someone may remember a book but be unable to find it. At another point, they may reach a document and need the page on which a passage appears. What would they need to verify before using the result?

Which part of the collection can we begin with?

The chosen enquiry will help identify which documents the application would need to process. Is there a set whose access conditions are already known? Before digitising it, each document will have to be situated within the collection and its relationship with the original established (primary physical document).

Where will the materials be processed?

Access conditions will help determine the path each document can follow. Which may use an external service? Which materials will have to be processed within the institution or remain in an isolated environment?



What work
can the
centre
sustain?

A small batch of documents would make it possible to calculate how long it takes to prepare them and review the OCR output. That experience would help assess the available equipment and the work that those responsible will be able to maintain after the pilot.

Which path do
we need to
be able to
reconstruct?

When the application finds a passage, can we return to the PDF and the page from which it came? If an error appears, will we know what happened during processing and which records depend on that page?

How far
should
this first
application
go?

An enquiry may be resolved through the catalogue or by searching within the text. On the other hand, some questions require passages to be gathered and worked with in a conversation with sources. The key question, then, is: which function does the enquiry chosen for the pilot actually need?

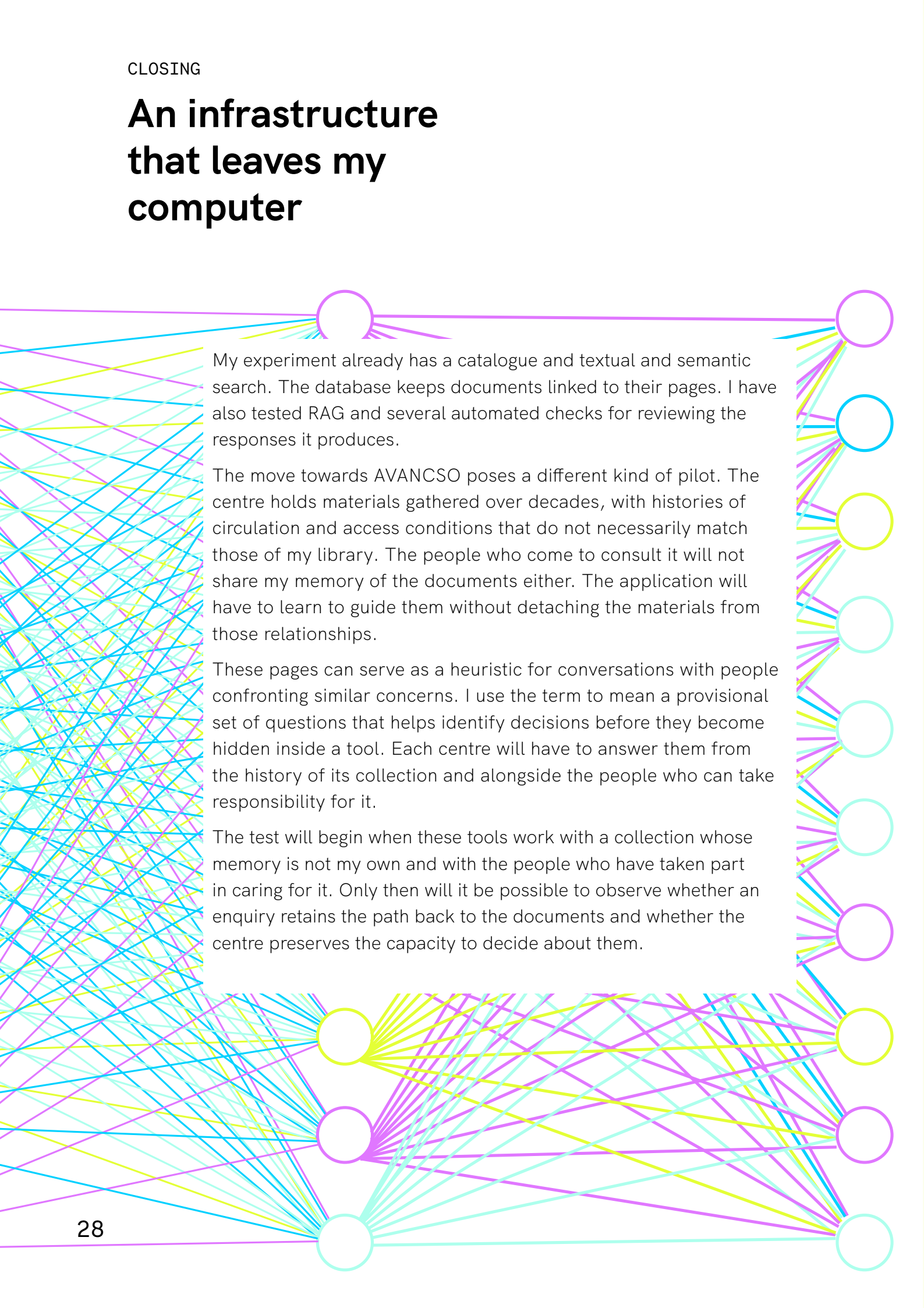
Which
functions
can wait?

The librarian-like assistant could resume a research process over several days or prepare an external document for intake. The pilot can retain these possibilities as a horizon without building them all in its first stage. In relation to the previous question, we must also ask: Which task do we need to test now, and which can be developed later?

Who will
decide, and
who will
be able to
continue the
work?

The centre will have to identify who may authorise a document's incorporation or approve a correction. Changes to the infrastructure will be recorded so that another person can reconstruct them. The test will be to ask that person to continue a task and observe which decisions still depend on the person who built the first version.

An infrastructure that leaves my computer



My experiment already has a catalogue and textual and semantic search. The database keeps documents linked to their pages. I have also tested RAG and several automated checks for reviewing the responses it produces.

The move towards AVANCSO poses a different kind of pilot. The centre holds materials gathered over decades, with histories of circulation and access conditions that do not necessarily match those of my library. The people who come to consult it will not share my memory of the documents either. The application will have to learn to guide them without detaching the materials from those relationships.

These pages can serve as a heuristic for conversations with people confronting similar concerns. I use the term to mean a provisional set of questions that helps identify decisions before they become hidden inside a tool. Each centre will have to answer them from the history of its collection and alongside the people who can take responsibility for it.

The test will begin when these tools work with a collection whose memory is not my own and with the people who have taken part in caring for it. Only then will it be possible to observe whether an enquiry retains the path back to the documents and whether the centre preserves the capacity to decide about them.

A step-by-step guide to the final stages

Summary of steps for setting up AI applications in documentation centres

PRINCIPLES FOR GETTING STARTED

Define metadata, access conditions, and traceability from the outset for both the original material and its derivatives (transcriptions, indices, and AI models); thus, should access rules change, you will be able to trace and intervene immediately across any existing copy.

Digitise relationships, not merely documents. The digital representation must preserve its relationship with the original, its provenance, and the conditions under which it is consulted.

Always preserve the original scan as a master copy: keep it separate from the working versions generated from it. In this way, you will always be able to generate new derivatives from the original master image.

Treat Optical Character Recognition (OCR) as an output that may contain errors. The extracted text must remain linked to the source image and page, whilst the processing record should allow for a new review when a problem emerges.

Ensure that any search or AI result can be verified by returning to the precise document and page of origin; once this traceability is established, tools such as RAG or semantic search may be incorporated as complementary pathways without relying on any single method.

Document for continuity

README files as a field journal. Record technical decisions and project milestones in README files to pass the “second-person test,” enabling future team members to maintain infrastructure continuity.

Check-list

Begin with an enquiry

Identify a specific research difficulty to determine the application's required function and delimit the subset of the collection to be processed.

Assess ethical issues and access conditions

Review archival permissions and establish the processing environment (external services, local processing, or isolated offline environments) to safeguard information and preserve institutional sovereignty.

Document preparation

Situate materials within the catalogue, scan physical documents, and apply Optical Character Recognition (OCR) to transform image pages into searchable text.

Data structure and traceability

Organise extracted text into JSONL or SQLite databases using stable content identifiers, ensuring every text fragment maintains a verifiable link back to its specific page and original PDF.

Retrieval, RAG, and automated checks

Combine deterministic textual matches (exact keyword/metadata search) with semantic search and Retrieval-Augmented Generation (RAG), employing automated tests to verify that generated responses remain grounded in the source passages.

Implement a dual-aspect modular interface (Librarian-like Assistant)

Research Tools: Catalogue & retrieval, annotated bibliography, conversation with sources, and research drawer.

Collection Care Tools: Document intake space and review tray for updating catalogue records and managing OCR corrections.

By: Dr Alejandro Flores-Aguilar

We thank the team at the AI Learning Labs: Andrés Vázquez, Kannika Thaimai, Leticia Coelho, Lucas Pretti, Patricio del Boca, Renata Ávila, and Solana Larsen.

Editor: Paz Peña

Illustrations and design: Constanza Figueroa Bustos

October 2026

Recommended citation style: Alejandro Flores-Aguilar. (2026). A library that supports research into a collection – Questions and decisions for developing AI applications in documentation centres from an anthropologist’s perspective. Open Knowledge Foundation.
<https://10.5281/zenodo.23067760>

Open Knowledge Foundation (OKFN) is a non-profit organisation registered in England and Wales with company number 05133759. okfn.org

Openness is our foundation.



We are grateful for the support of the [Patrick J. McGovern Foundation](#). Find out more about their funding programmes [here](#).



This publication is licensed under a Creative Commons Attribution 4.0 International Licence ([CC-BY 4.0](https://creativecommons.org/licenses/by/4.0/)).



AI

Learning

Labs

#AllLearningLabs

Catalysing and developing replicable methods to help organisations build AI skills, use AI responsibly and develop their own AI projects. Further information: okfn.org/ai-learning-labs



Get involved

Run a pilot: Interested in testing this? Reach out to us at info@okfn.org

Join the conversation: Have lessons or questions from your own work? Connect with other practitioners in our [Open Knowledge Forum](https://openknowledgeforum.org).



Open Knowledge